DTIC
ELECTE
JAN 03 1995
S D
B

19941228 020

Computer-Aided Breast Cancer Diagnosis

THESIS

Catherine Mary Kocur
Captain, USAF

AFIT/GSO/ENS/94D-03

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GSO/ENG/94D-03

Computer-Aided Breast Cancer Diagnosis

THESIS

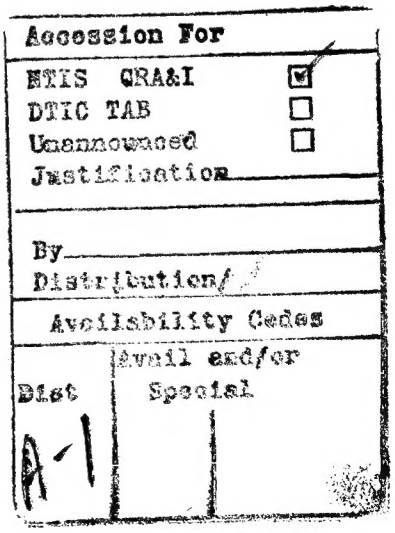
Catherine Mary Kocur
Captain, USAF

AFIT/GSO/ENS/94D-03

Approved for public release; distribution unlimited

DTIC QUALITY INSPECTED 3

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U. S. Government.



AFIT/GSO/ENG/94D-03

COMPUTER-AIDED BREAST CANCER DIAGNOSIS

THESIS

Presented to the Faculty of the Graduate School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Space Operations

Catherine M. Kocur, B.S.

Captain, USAF

DECEMBER 1994

Approved for public release; distribution unlimited

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor, Doctor Steve Rogers, for his constant support and inspiration. His enthusiasm is contagious. My gratitude also goes to my committee members, Lt Col Kenneth Bauer and Captain Dennis Ruck, for their comments and guidance. I will always be grateful to Captain Jeff Hoffmeister for the education and scrutiny he provided throughout this endeavor. Additionally, I would like to recognize Shireen Braner of Grandview/SouthView Hospitals, without whose cooperation and assistance this endeavor would never have been able to get off the ground.

Recognition must also go to the computer lab gang -- Lem, Chris, Terry, and all the rest -- for their support and contributions. Dave and Dan I am sure will appreciate having their disk space back, and I am very happy to include them here with heartfelt thanks for helping me over those numerous stumbling blocks.

To the Deiverts, my surrogate family, and the Space Operations Bunch, thanks for helping me keep my insanity. And to Monte, I do not know how you managed to provide so much support and motivation over such long distances, but for that I am extremely grateful.

Table of Contents

Page

Acknowledgments ii

Table of Contents iii

List of Figures vii

List of Tables ix

Abstract x

I. Introduction 1-1

1.1 Cancer Facts 1-1

1.2 Computer-Aided Diagnosis 1-3

1.3 Background 1-3

1.4 Statement of Problem 1-5

1.5 Scope 1-5

1.6 Thesis Overview 1-6

II. Background 2-1

2.1 Breast Cancer Basics 2-1

2.2 Risk Factors 2-3

2.3 Mammographic Acquisition Technologies 2-4

2.3.1 Film 2-5

2.3.2 Digitized Film 2-7

2.3.3 Digital 2-7

Page

2.3.3.1 Stereotactic Imaging 2-7

2.4 Database Selection and Reporting 2-8

2.4.1 ROC Curve 2-9

2.5 Segmentation Techniques 2-11

2.6 Feature Extraction 2-13

2.6.1 Microcalcification Detection 2-14

2.6.2 Mass Detection 2-17

2.6.3 Parenchymal Pattern Detection 2-18

2.7 Feature Selection 2-19

2.8 Classification Methods 2-20

2.8.1 Bayesian Classifiers 2-20

2.8.1.1 Gaussian Classifier 2-21

2.8.1.2 K-Nearest Neighbor 2-21

2.8.2 Artificial Neural Networks 2-21

2.8.3 Post Multiplying the Classifier Output 2-22

2.9 Summary 2-23

III. Methodology 3-1

3.1 Angular Second Moment 3-1

3.1.1 Gray-level Histograms 3-1

3.1.2 ASM Calculation 3-3

3.2 Karhunen-Loeve Transform 3-4

Page

3.2.1 Eigenvalues/Eigenvectors 3-5

3.2.2 Eigenmass Method 3-5

3.3 Wavelets 3-7

3.4 Summary 3-9

IV. Database and Analysis 4-1

4.1 Database 4-1

4.1.1 Digital 4-1

4.1.2 Digitized 4-2

4.1.2.1 File Format 4-4

4.1.2.2 ROI Selection 4-4

4.1.3 Other Sources 4-6

4.2 Image Display 4-6

4.2.1 Khoros 4-7

4.2.2 The Mammogram Image Processing Program 4-7

4.3 Classification 4-8

4.4 Individual Feature Performance 4-9

4.4.1 Angular Second Moment 4-9

4.4.2 Eigenmass 4-9

4.4.3 Wavelets 4-9

4.5 Feature Selection 4-11

4.5.1 Probability of Error 4-11

Page

4.5.2 Steppe Algorithm 4-14

4.5.3 Lee and Langrebe Theory 4-17

4.6 Summary 4-19

V. Results and Conclusions 5-1

5.1 Summary 5-1

5.2 Conclusions 5-2

5.3 Recommended Future Research 5-3

Appendix A. Terms A-1

Appendix B.1 Stereotactic Mammographic Images Information B-1

Appendix B.2 ROI Details B-3

Appendix B.3 Mammogram Screening/Pathology Reports B-4

Appendix B.4 LORAD System Specifications B-14

Appendix C. Cincinnati Database Digital C-1

Appendix D. Other Databases D-1

Appendix E. Khoros E-1

Appendix F. Matlab Code F-1

Bibliography BIB-1

Vita V-1

List of Figures

Figure — Page

2.1 Effects of Contrast, Sharpness and Noise in the Portrayal of Fine Mammographic Detail 2-5

2.2 Schematic Top View of a Stereotactic Localization System 2-8

2.3 Probability Distribution for ROC Analysis 2-10

2.4 Example ROC Curve 2-11

2.5 Probability of Error Determination 2-19

3.1 Gray-Level Values of an Example Image 3-1

3.2 First-Order Gray-Level Histogram 3-1

3.3 Mammographic Image with Microcalcifications 3-2

3.4 Mammographic Image Histogram 3-2

3.5 Second-Order Gray-Level Histogram 3-3

3.6 Eigenvectors 3-4

3.7 Wavelet Decomposition Process 3-8

3.8 Wavelet Decomposition for Mammographic ROI 3-8

4.1 Digital Mammogram Containing a Mass Lesion 4-1

4.2 Digital Mammogram Containing Microcalcifications 4-2

4.3 Digitized Mammogram with Microcalcifications 4-3

4.4 Sub-Region Image 4-3

4.5 Region of Interest Image 4-5

Page

4.6 Khoros Workspace 4-8

4.6 Low Probability of Error Feature 4-12

4.7 High Probability of Error Feature 4-13

4.8 Scree Plot of Saliency Means 4-15

4.9 Percent Error Change During Model Reduction 4-16

4.10 Feature Combination for the Largest Eigenvalue Vector 4-17

4.11 Lee and Langrebe Eigenvectors 4-18

4.12 Lee and Langrebe Classification Results 4-19

List of Tables

Table		Page
1.1	Chances of Developing Breast Cancer	1-2
2.1	Criteria Used for Diagnosis	2-1
2.2	Risk Factors for Breast Cancer in Females	2-4
2.3	Average PEC Values for All Features	2-15
2.4	Results of Cluster Analysis for Selection of Best Feature	2-16
2.5	Frequency-of-Occurrence List of Features	2-16
4.1	Classification Performance for ASM	4-9
4.2	Classification Performance for Eigenmass	4-9
4.3	Classification Performance for Wavelets	4-10
4.4	Steppe Saliency Means and Ranks	4-14
4.5	Steppe Model Selection	4-16
4.6	Feature Selection Summary	4-19

ABSTRACT

This research advances computer-aided breast cancer diagnosis. More than 50 million women over the age of 40 are currently at risk from this disease in the United States. Computer-aided diagnosis is offered as a *second opinion* to radiologists to aid in decreasing the number of false readings of mammograms. This automated tool is designed to enhance detection and classification.

New feature extraction methods are presented that provide increased classification power. Angular second moment, a second-order gray-level histogram statistic, provides baseline accuracy. Two novel extraction methods, *eigenmass* and wavelets, are introduced to the field. Based on the Karhunen-Loeve Transform, *eigenmass* features are developed using eigenvectors to alter the data set into new coefficients. Wavelets, previously only exploited for their segmentation benefits, are explored as features for classification. Daubechies_4, Daubechies_20, and biorthogonal wavelets are each investigated. Applied to 94 *difficult-to-diagnosis* digitized microcalcification cases, performance is 74 percent correct classifications.

Feature selection techniques are presented which further improve performance. Statistical analysis, neural and decision boundary-based methods are implemented, compared, and validated to isolate and remove useless features. The contribution from this analysis is an increase to 88 percent correct classification in system performance.

COMPUTER-AIDED BREAST CANCER DIAGNOSIS

I. Introduction

Breast cancer detection and diagnosis continue to plague women and daunt researchers worldwide. Although technological progress has increased a woman's chance for survival, continuing developments in digital image processing can further reduce morbidity and mortality rates from this disease. The Air Force Institute of Technology (AFIT) has worked smart weapons and other related military image processing problems for thirty years. It is the goal of this thesis to transition this technology to solve medical problems.

1.1 Cancer Facts

The National Cancer Institute (NCI) estimates that 182,000 women will be newly diagnosed with breast cancer this year (31:1) -- one every three minutes, with approximately 46,000 deaths from the disease. More than 50 million women over the age of 40 are currently at risk in the United States. With a frequency of six cancers per 1,000 women screened, and assuming a 50 percent screening rate of those at risk, approximately 150,000 new individuals will be diagnosed each year (12:540). It is also known that 55 percent of these new cancers will be localized to the breast, while 45 percent are metastatic (having already spread).

Today, a woman has a one in eight chance of developing breast cancer in her lifetime. Concern increases for women over age 50, since 75 percent of all breast cancers occur in females from this group. Table 1.1 illustrates the dramatic affect age has on a woman's chance for developing breast carcinoma (53:54).

Table 1.1 Chances of Developing Breast Cancer

BY AGE 25: ONE IN 19,608	BY AGE 60: ONE IN 24
BY AGE 30: ONE IN 2,525	BY AGE 65: ONE IN 17
BY AGE 35: ONE IN 622	BY AGE 70: ONE IN 14
BY AGE 40: ONE IN 217	BY AGE 75: ONE IN 11
BY AGE 45: ONE IN 93	BY AGE 80: ONE IN 10
BY AGE 50: ONE IN 50	BY AGE 85: ONE IN 9
BY AGE 55: ONE IN 33	IN A LIFETIME: ONE IN 8

These figures represent lifetime risk. A different view of these statistics was shown by the American Cancer Society which reported that a white woman's chance of developing breast cancer from birth to 50 was two percent; from birth to 70, was about 6 percent. Note, however, the lifetime risk factor of one in eight is based on a 95 and up age group. This can be misleading since the average life expectancy of women in the United States is 79 (53:54). For most, the one in ten chance of developing breast cancer is a more accurate estimate.

Although mammography is currently the best method for breast cancer screening, 10 percent of breast cancers -- even lumps you can feel -- do not show up on x-rays (31:4). Silverberg et al (48) states mammography reduces the mortality from breast cancer by 30 to 35 percent. However, 10 to 30 percent of negative readings are later proven to be from patients with breast cancer; two-thirds of which are evident in the mammogram in retrospect (17:283). These false-negatives can be attributed to several factors including poor image quality, radiologist fatigue, and just plain oversight. To counter some of these factors it has been suggested that a *second opinion* may be the

answer (17:283). With the already high volume of mammograms a radiologist must view, using another radiologist for a second opinion would further deteriorate the screening process. With the system already backlogged, an increase in the number of mammograms a radiologist must review would simply increase fatigue even further. It is the hope that computer-aided diagnosis (CAD) will provide radiologists with this assistance.

1.2 Computer-Aided Diagnosis

Computer-aided diagnosis offers an automated tool to enhance detection and classification of breast cancer. It can be defined as a diagnosis made with input from computer analysis of radiographic images.

Enhanced detection is the first objective of a CAD system and offers an increase in the effectiveness and efficiency of mammographic screening. Improved classification, the second objective, assures more accurate and possibly earlier diagnosis for the patient. False-positives, diagnosing an abnormality as malignant when it is actually benign, causes unnecessary surgery and trauma. A false-negative reading calls a malignant area benign. This misdiagnosis could prove fatal. Therefore, with improved classification, morbidity, mortality, and number of biopsies could all be decreased.

The ultimate test of a CAD system is to compare the performance of radiologists alone to radiologists having been provided with a computer opinion.

1.3 Background

Breast cancer research is vast and a continuing topic of concern. Mammographic acquisition technology, segmentation, feature extraction, and classification are all areas that need to be addressed when delving into breast cancer detection and classification.

Mammographic imaging developed from early x-ray techniques to the use of xeromammography which follows the same principles as a copy machine of putting ink on paper. X-ray film then again became prominent due to technological improvements. For computer manipulation purposes, the field has progressed to digitization of these films. More recently, procedures for direct digital acquisition have entered the field. Another new development has been the introduction of stereotactic breast imaging and biopsy. This method provides images of a selected region of interest with the same historical mammographic view along with several images taken off-axis. Studies have been done using and comparing each of these acquisition technologies (7, 56).

Preprocessing and segmentation have posed unique image processing problems. Common techniques, such as contrast variations and gray-level histogram manipulation, have not yielded promising results. As a result, several researchers have combined manual identification with automated classification (7,44).

Mass lesions and microcalcifications, explained further in Chapter II, are the two types of cancer found in the breast. Since each abnormality is unique, most researchers elect to concentrate their studies on one type. Furthermore, extraction methods have involved both computer-recognized and human-identified features (9,17,33,57).

Classification has called upon various techniques, such as Gaussian classifiers, probability of error methods, and artificial neural networks. Some researchers select a classifier or use several to compare results (9,17). Display of each system's performance poses yet another unique problem and is identified further in Chapter 2.

In reviewing the background literature, is was determined that research here at AFIT could offer great contributions in automated feature extraction, feature saliency, and artificial neural networks.

1.4 Statement of Problem

The ultimate objective of this study is to establish an efficient computer-aided diagnosis system that will extract a conventional second-order gray-level histogram feature, a Karhunen-Loeve transform feature, and wavelet features that have been proven in military research (49). The resulting classification, benign or malignant, will be provided to the radiologist.

1.5 Scope

The initiatives covered in this study use both digital and digitized mammograms. Building a database of digital mammograms, which provide better resolution, continues to be an ongoing endeavor. However, the number of images in this database is still too small to obtain reliable results when testing the performance of our CAD system. Therefore, digitized "hard-to-diagnose" mammograms were obtained (9) and are used to train and evaluate the system.

Segmentation is performed on the raw digital images with a physician, Doctor Jeff Hoffmeister, manually selecting a sub-region which contains the area of interest. Similar sub-regions, identified using binary masks (9:1380), were provided with the digitized database. For this research endeavor, it is assumed that segmentation was performed correctly and those regions with abnormalities were properly identified.

Feature extraction methods explore the use of angular second moment (ASM), *eigenmasses*, and wavelets; all of which are described in detail in Chapter 3. ASM, is a second-order gray-level histogram statistic, whereas eigenmasses are developed using a Karhunen-Loeve Transform (52). Wavelet feature extraction investigates the different contributions of Daubechies_4, Daubechies_20, and biorthogonal wavelets.

Feature saliency and selection are investigated using several methods. Probability of error statistics provide a first glance at feature saliency. Steppe (50,51) incorporates techniques developed by Priddy et al (37), Ruck et al(41), and Tarr (54). Lee and Langrebe (30) saliency is also investigated. Applying these techniques distinguishes redundant or valueless features.

Selected features are then entered into LNKnet, a pattern analysis package, for classification. Several schemes are available and explored, with emphasis on artificial neural networks, using this package.

1.6 Thesis Overview

With the cancer statistics and introduction provided in Chapter I as a foundation, Chapter II examines background information and related research. Types of mammography, database issues, and examinations of segmentation, feature extraction, feature saliency, and classification are addressed. Chapter III discusses the specific methodology used for this study. Data description and analysis are presented in Chapter IV, with the findings of this research and conclusions given in Chapter V.

2.1 Breast Cancer Basics

Breasts are made up of fatty tissue, supporting ligaments, connective tissue, milk-producing glands, and ducts. *Screening* mammograms, those done on individuals with no symptoms or history, are suggested for women over age thirty-five. *Diagnostic* mammograms are done when symptoms, history, or physical findings advise.

Table 2.1 Criteria Used for Diagnosis (23)

MASSES	
BENIGN	**MALIGNANT**
Round	Spiculated
Smoothly defined margins	ill-defined margins lobulated margin
Do not usually cause distortion of normal breast tissue	Areas of asymmetrical increased density or distortion of normal breast architecture
Multiple or present in both breasts	Solitary

MICROCALCIFICATIONS	
BENIGN	**MALIGNANT**
	0.1-0.5 mm (can be up to 1mm) in diamet
Less than 5 in 1 ml volume of breast tissue	5 or more in 1 ml volume of breast tissue
More regular, smooth shape	Irregular in shape with pointed edges
Larger and thicker	Smaller and thinner
Diffusely scattered in both breats	

SECONDARY FEATURES	
BENIGN	**MALIGNANT**
"Popcorn" appearance of a calcifying involuting fibroadenoma	Asymmetric prominent ducts
	Skin or nipple thickening
	Retraction
	Asymmetrically enlarged veins
	Axillary lymph node involvement

Cancer may appear in the breast as a mass or small worm-like deposits of calcium called calcifications (normally microcalcifications). One thing to note is the fact that benign calcifications are a normal occurrence in the breast. A preliminary diagnosis will be made from the mammogram using guidelines similar to those shown in Table 2.1. Wu et al (57:82) also provides an extensive list of diagnostic characteristics. These characteristics make a lesion (mass or microcalcifications) *more likely* to be benign or malignant. Most lesions will have characteristics from both the benign and malignant lists and the radiologist must weight the importance of the features to classify the lesion as more likely malignant or benign.

Once a suspicious abnormality is detected, a biopsy is normally performed to diagnose the lesion as malignant or benign. A case sent for biopsy has a 30 percent chance of proving malignant (53:58).

> The least invasive technique is fine-needle-aspiration biopsy, in which a very thin needle is inserted into the suspicious area and a few cells are drawn out. Core-needle biopsy is a similar technique that removes a sample of tissue, not just a few cells. In stereotactic needle biopsy, a computer calculates a three-dimensional view of the breast, using mammography, and inserts the needle into the suspicious area. The technique is useful when a suspicious area can be seen by mammography but not felt. (53:60)

The biopsy sample is forwarded to pathology where gross (visible to the naked eye) and microscopic examinations are performed. A pathologic diagnosis is rendered. Appendix B contains pathology results for the images that make up the digital database used in this study.

The diagnosis will include stages to describe the severity of the cancer. A higher stage has a higher morbidity and mortality rate. In situ carcinoma is a pathological diagnosis, not a clinical stage as are Stages I through IV. In situ carcinoma denotes that the cancer is noninfiltrating. In other words, the cancerous cells are still contained within the confines of the region in which they arose. For example, in an in situ ductal breast

carcinoma, the cancerous cells are still within the ductal epithelium. In situ carcinoma can almost be considered Stage 0 or a pre-cancerous condition. It is felt that in situ carcinoma will progress to an infiltrating cancer if not treated. Stage I identifies a small tumor (less than two centimeters in diameter) that has not spread beyond the breast. Stage II is a larger tumor (less than five centimeters in diameter) that has spread to nearby lymph nodes, whereas Stage III is a tumor larger than five centimeters that has spread to other than nearby nodes. Stage IV is reserved for cancer that has spread elsewhere in the body (23,53).

Appendix A lists terms and definitions referred to throughout this study. Mayo (31) and Tanne (53) are suggested reading on breast cancer facts and statistics.

2.2 Risk Factors

Risk factors are also being used to ascertain the possibility an individual develops breast cancer. Factors such as family history, environment, and social status may all play key roles. It is interesting to note, however, that the American Cancer Society estimates 75 percent of breast cancers occur in women with no high risk factors. Table 2.2 is an excerpt from the list provided by Tanne (53:58) that identifies common risk factors.

Table 2.2 Risk Factors for Breast Cancer in Females

Significantly higher risk ***	Age Country of Birth Family medical history	50 or older North America, northern Europe Mother or sister with history of breast cancer
Moderately higher risk **	Socioeconomic status Age at first pregnancy Personal medical history Family medical history	Upper 30 or older Previous cancer in one breast Benign tumor (fibroadenoma) Mother or sister with history of breast cancer
Slightly higher risk *	Martial status Place of residence Race Duration of estrogen exposure Number of pregnancies Weight Personal medical history	Never married Urban; Northern United States Caucasian women age 45 or older African-American women younger than age 40 Menopause after age 55 Menstruation before age 11 None Obesity after menopause Previous endometrial or ovarian caner

The risk column illustrates a difference between women at high and low risk. For example, *** means that this risk is four times greater for an individual that falls into this category. Information like the fact that twenty percent of women with breast cancer have a first or second (grandmother) degree relative with the disease could be used in a CAD system. One possible approach would be to use this knowledge to develop a separate CAD system that predicts an individual's chance of developing cancer. Another method would be use the risk factors to customize the CAD system. For example, weights could be assigned to each risk factor and used to control the sensitivity that the system uses in its diagnosis. Those with higher risk factors would be scrutinized more closely (threshold levels set lower). Development of a CAD system that exploits risk factors is identified in Chapter V as an area for future research.

2.3 Mammographic Acquisition Technologies

X-ray film is the accepted medium for mammographic images. For computer display and manipulation these images are commonly digitized. Most recently, digital

breast imaging has entered the scene with stereotactic system utilizing digital acquisition to produce multiple views of the breast to calculate lesion depth for biopsy.

2.3.1 Film. Resolution for x-ray film varies between 15 and 22 line pairs per millimeter (lp/mm). The value 22 lp/mm is used as a typical representation and equates to 25 microns resolution. (20)

Vyborny and Schmidt (56) discuss how the detection of early breast carcinoma is effected by the technical quality of film mammograms. X-ray processing and recording, as well as film processing, all effect the visualization of fine mammographic detail. Vyborny and Schmidt further divide the technical quality of mammographic images into two categories: "(a) those commonly associated with the physics of the x-ray imaging process and (b) those associated with performance of the examination by the technologist." The physics-related topics of contrast, sharpness, and noise and their effects are discussed in detail.

Before taking a closer look at each of these factors, it must be understood that these components are not readily separable. As the simple example in Figure 2.1 shows, unsharpness will blur and widen the area depicted by contrast alone. The addition of noise causes further deterioration in the ability to identify the region.

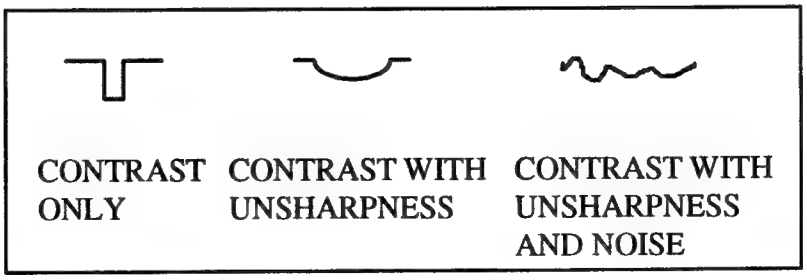


Figure 2.1 Effects of Contrast, Sharpness and Noise in the Portrayal of Fine Mammographic Detail (46, 102)

High contrast, a large difference between a target and its background, increases the possibility of detection of subtle mammographic abnormalities and, therefore, has been an area of intense research. One reason that the improvement of contrast sensitivity is so difficult, is the problem that simple forms of noise effect both the signal and the noise equally.

The overall contrast of an image is a result of subject contrast and recording-system contrast. Factors affecting subject contrast are x-ray attenuation, x-ray scattering, and the spectrum of x-ray used. Breast compression increases image contrast by decreasing the amount of scattered radiation. The properties of the mammographic film and developing manner affect the recording-system contrast.

Unsharpness hampers fine mammographic abnormality detection by causing blurring. The different components of unsharpness are geometric or focal-spot unsharpness, motion blurring, and receptor unsharpness. An example of the adverse affect unsharpness has on detection is seen with fine spiculations. Spiculations radiate from a single point, forming a star-like pattern. This feature is often used by radiologist to identify suspect regions. If the spatial resolution of the recording process is poor, unsharpness is introduced and the spiculations may become indistinguishable.

Since noise is too closely related to the effects of contrast and sharpness it has been difficult to isolate the lone affect of noise. However, the "reduced viewing distance used in mammography, as well as the use of optical magnification, accentuates the perception of noise (56:106)."

2.3.2 Digitized Film. Digitization of film mammograms provides the luxury of being able to manipulate the needed image with computer techniques. Chitre et al (9:1380) details their digitization method and reports the specifications of the resulting images.

Some consideration must be taken when working with digitized images. The digitization process can vary in terms of pixel size and number of quantization levels. Therefore, quality of the resulting image is dependent on technique and capability of the system used for digitization. Computer-aided diagnostic schemes may possibly take into account the size of the lesion to pick pixel size. Detection of microcalcifications requires a much smaller pixel size (approximately 100 microns (47)) than that needed for mass detection. Classification schemes require even higher resolution for finer mammographic detail.

2.3.3 Digital. Digital imaging has yet to be widely accepted in this field. Studies, such as the one referenced in Appendix D and currently taking place at Georgetown University, provide in-depth examinations on the improved capabilities digital images provide, comparison of these images to film, and test the performance of radiologists using this new medium.

For digital mammograms, resolution of "100 microns may be acceptable, however, 50 microns represents a reasonable and achievable goal (59)."

2.3.3.1 Stereotactic Imaging. Stereotactic breast imaging takes digital acquisition one step further and incorporates multiple images to obtain depth information on the lesion. In this non-surgical procedure, core biopsy specimens are obtained with a 14-gauge cutting needle. The patient is place in the prone position on a table with the breast protruding through a hole. The x-ray tube, compression device, image receptor,

and biopsy device are located under the table. Stereotactic views are used to guide needle placement and a sample of tissue is extracted.

With only a small amount of local anesthetic and a small incision at the needle entry point, the entire process takes only half an hour, costs one-third that of surgical excisional biopsies, implies minimal risk, and causes no residual scarring (22:257).

Stereotactic localization uses two radiographic views to determine site of the abnormality in three spatial dimensions (horizontal, vertical, and distance into the breast). The two views differ in horizontal position, as shown in Figure 2.2, with the beam commonly directed +15 degrees and -15 degrees relative to a line perpendicular to the image receptor (22:259). Although screen film is still available, in a stereotactic system the advantage comes from using the digital image receptors and having the image displayed in 10 to 20 seconds.

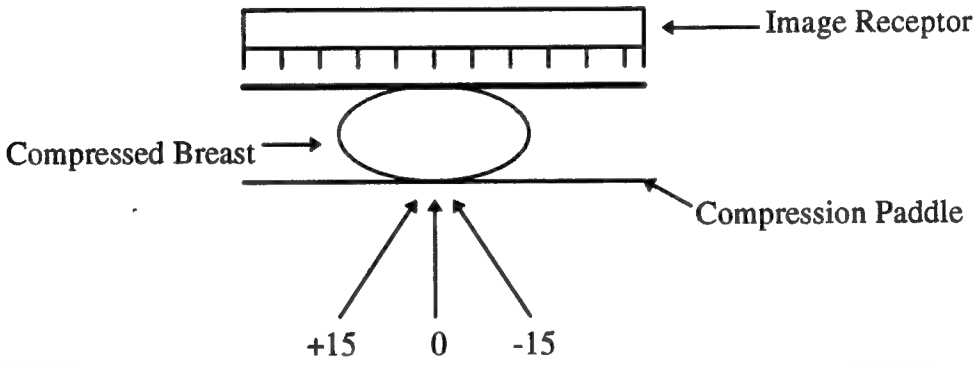


Figure 2.2 Schematic Top View of a Stereotactic Localization System

2.4 Database Selection and Reporting

Case selection is of utmost importance in evaluating the results of a CAD system. The type of data chosen greatly impacts the performance of the classifier used in the

system. Obviously, those studies which use difficult-to-diagnose cases cannot be compared to a system, no matter how similar the detection and classification methods are, with a scheme that employs easy-to-diagnosis cases. Another problem associated with each investigator using their own database is that it is difficult to predict how their system would perform under actual clinical use.

Nishikawa et al (35) conducted a study that shows the sensitivity of a test CAD scheme varied between 26 and 100 percent (with a false-positive rate of 1 per image) depending upon the cases selected for the database. They identified that a change of only 20 percent of the cases resulted in a 15 to 25 percent change in sensitivity. Nishikawa suggests several long-term solutions: exchanging databases (technical standardization would be required), developing a common database, or establishing a measure of image quality and case specification. In the interim, it is suggested that all investigators report the method used in selecting cases and characterize their database with histograms or the mean and standard deviation of image features.

2.4.1 ROC Curve.

The free-response receiver operating characteristic curve is becoming the standard method for graphically reporting the performance of CAD systems. Previously, only the percentage of correct decisions was reported. This performance measure, known as *diagnostic accuracy*, had obvious limitations (32:720). The most apparent is the fact that if a suspicious lesion has only a 30 percent chance of proving malignant, then by classifying all lesions as benign the diagnostic accuracy would already be 70 percent correct. Another shortcoming is realized in that no information is being provided on the relative frequencies of false-positive and false-negative errors.

The solution has been to report system performance in terms of trade-off values. The most common pairs are sensitivity versus specificity, or the corresponding true-positive fraction (TPF) and false-positive fraction (FPF).

$$\text{SENSITIVITY (TPF)} = \frac{\text{NUMBER OF CORRECTLY CLASSIFIED MALIGNANT CASES}}{\text{TOTAL NUMBER OF MALIGNANT CASES}}$$

$$\text{SPECIFICITY (FPF)} = \frac{\text{NUMBER OF INCORRECTLY CLASSIFIED BENIGN CASES}}{\text{TOTAL NUMBER OF BENIGN CASES}}$$

The ROC curve is derived from the discrimination capacity of the system, which is dependent upon the separation or overlap in the probability distributions of the feature being investigated. As seen in Figure 2.3, at a specific threshold there will be a certain percent correct for malignant and percentage of false alarms.

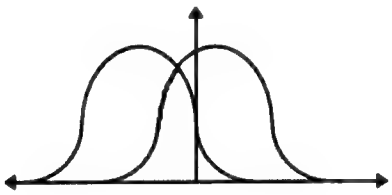


Figure 2.3 Probability Distribution for ROC Analysis

Using this threshold, the TPF and FPF can be determined and a single point established on the ROC curve. By varying the confidence threshold, the ROC curve represents all trade-offs achievable by the system.

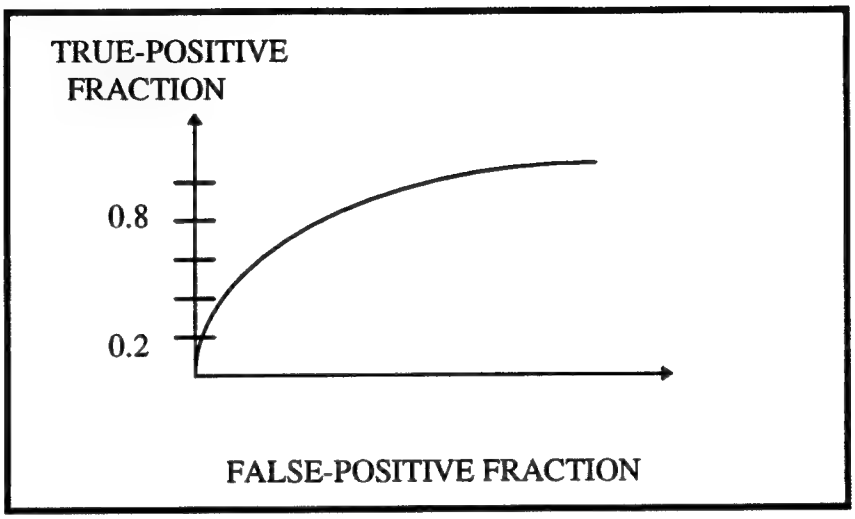


Figure 2.4 Example ROC Curve

The higher the curve is in the upper left corner, the better the system is at correctly identifying cancers. Figure 2.4 illustrates that this pictorial view gives the radiologist a quick reference of how reliable the diagnosis is that he/she is being fed. Another common method for reporting the same information is to simply provide the area under the curve.

2.5 Segmentation Techniques

Segmentation can be defined as identifying a region or regions of interest, commonly referred to as ROIs, in an image. These areas are normally groupings of similar properties and are located by enhancing or de-emphasizing certain characteristics of the raw image. Consideration must be given to the characteristic chosen. For example, methods employing contrast may be of use to investigators trying to locate tanks in aerial

photography, where it may sometimes be of little use in discerning abnormalities from normal tissue in medical images. The final objective of this step in image processing is to reduce the number of pixels that the following steps must consider.

Sarwal et al (44) uses a contrast enhancement function to modify the contrast distribution of an image. Region of interest segmentation is then performed using a region growing technique that selects high intensity pixels in this new contrast stretched image. The ROI is contrast stretched and further grown to result in microcalcification segmentation.

Chan (7) applies a difference-image method to identify microcalcifications. The image is spatially filtered twice: first increasing the signal-to-noise ratio, then reducing this signal-to-noise ratio. The difference between the two results in an image with background density variation removed -- a similar method to the technique developed by Harrup (19) that uses a hit/miss morphological transform. In the Chan study, global and local thresholding is then used to segment out the microcalcifications. This study was taken to the classification stage with only those clusters with larger than a certain number of calcifications being retained. Performance for the system was reported as 85 percent sensitivity with 2.5 false-positive per image.

In recent years, wavelet transforms have been applied. The literature reviewed showed most researchers concentrate on wavelet contributions in preprocessing (3,5,8,28,29,60). Wavelets have been applied for contrast enhancement, edge identification, and signal-to-noise manipulation to improve visualization with the hope of earlier detection.

2.6 Feature Extraction

In image processing, feature extraction involves processing raw data to develop a set of measurements that represent a ROI and can be used for classification. No information can be gained during this procedure. In theory, processing raw data would provide the best results in the classification stage. But in almost all cases, cancer detection being one of them, this option is not feasible. The dimensionality alone prevents us from using raw data in the classifier.

The objective of feature extraction is to develop a set of features, each obtained from measurements taken from the image, to represent the region of interest for classification. To identify breast carcinoma a wide range of features have been applied, such as asymmetry between the left and right breast, intensity, spiculation, and texture -- just to name a few. This study investigates second-order gray-level histograms, a Karhunen-Loeve Transformation, and wavelets.

Wu and colleagues (57) conducted a study in which 14 expert radiologist manually extracted 43 features from the mammograms. These features where broken into 3 groups: features associated with masses (shape, size, spiculation, pattern), characteristics related to microcalcifications (number, shape, uniformity, distribution), and secondary abnormalities (parenchymal distortion, skin thickening). A complete listing on the features the radiologist were looking for is provided (57:82) as part of their study. A statistical feature selection method was employed and the dimensionality reduced to 14 features. Those features retained are also provided (57:86). Classification was performed using a three layer, feedforward neural network trained using backpropagation to discriminate 133 benign and malignant cases. Performance results were displayed using ROC curve and showed that with an area of 0.95 under this curve, the neural network performed at a level

higher than averages achieved by attending radiologists, residents, and experienced mammographers.

Although not specifically applied to breast cancer detection, Pietikainen (36) uses Laws texture features that measure the average degree of likeness between pixel neighbors identified by applying a set of masks. It is stated that these features perform better than methods based on pairs of pixels. Each mask and its effect are described.

Miller (33) also suggests that texture, rather than intensity, may be a better discriminator. This examination employed both Serra and Laws texture measures to identify asymmetry. Using a Bayesian classifier with a leave-one-out method, an 80 percent correct classification rate was reported for their database of 40 images.

2.6.1 Microcalcification Detection. Breast cancer is detected 21 percent of the time due to the identification of microcalcifications (12:540). However, 30 to 50 percent of all cancers detected by mammography contain microcalcifications (17:284), whether alone or with a mass. Of the cancers associated with calcifications, 71 percent are minimal (invasive cancer less than five millimeters in size or wholly intraductal) (12:540). Size for malignant microcalcifications commonly varies between 100 and 300 microns (47).

Recent work by Chitre et al (9,14,13,10) developed the idea of using image structure features, which are derived from the second order gray-level histograms, to identify and diagnose microcalcifications. Using 95 "difficult-to-diagnose" cases, 57 proven benign and 38 proven malignant, these mammograms were digitized to a resolution of 160 microns per pixel. The image structural features where introduced to a feedforward backpropagation neural network for classification. The best result identified, was a sensitivity rate of 73 percent with a false-positive rate of 35 percent.

After identifying that general region-growing and thresholding base segmentation failed to be accurate, Chitre manually extracted a region of interest using a binary mask. Further research was used to prove that the binary mask is size indifferent. The ROI was then normalized to gray-levels of 0 to 255.

The first-order gray-level histogram is defined as the distribution of the probability of occurrence of a gray-level in an image. The second-order gray-level histogram is the probability distribution of a pair of gray-levels separated by a given distance. These histograms are described in detail in Chapter III. Second-order histograms provide the advantage of not requiring binary segmentation which has produced poor results. Chitre selected ten image structural features for further analysis. Table 2.3 lists the chosen features.

Feature selection was done using a combination of Probability of Error Classification (PEC), which is presented in the feature selection section of this chapter, and K-Means clustering. The conditional mean and variance was calculated for each feature to investigate their Gaussian probability density functions (pdf). As shown below, Chitre rank ordered the features by placing the features with the lowest possibility of error at the top.

Table 2.3 Average PEC Values for All Features(9:1391)

FEATURE IDENTIFICATION	AVERAGE PEC VALUE
Angular Second Moment	0.1846
Angular Second Moment: Diff.	0.3149
Deviation	0.3469
Contrast	0.372
Mean	0.3963
Entropy: Diff.	0.4057
Mean: Diff.	0.4099
Correlation	0.4114
Entropy	0.4161
Inverse Difference Moment	0.4602

The K-Means Method used clustering techniques to develop a list of cluster numbers and feature sets. The features were then organized in a frequency of occurrence (FOC) table. Table 2.4 and Table 2.5 show the results of this investigation.

Table 2.4 Results of Cluster Analysis for Selection of Best Feature(9:1395)

No. of Clusters	Set of Features	Performance
4	0,1,2,5,7	56.85
4	0,2,3,4,5,6,7	73.16
5	0,2,3,6,7,8	63.79
5	0,1,2,3,6,7,8	68.23
6	1,2,3,8,9	67.72
6	0,2,3,6,7,8,9	72
7	1,2,3,5,6,7,8	73.96
7	0,1,2,3,5,6,8,9	74.92

Table 2.5 Frequency-of-Occurrence List of Features(9:1395)

FEATURE	FREQUENCY OF OCCURRENCE
Deviation	8
Entropy: Diff.	7
Contrast	6
Entropy	6
ASM	6
Inverse Diff. Moment	6
Correlation	5
Mean	4
ASM: Diff.	3
Mean: Diff.	1

Two sets of features were selected. Feature set one used the three features common to the top five from both the PEC and FOC tables. Angular Second Moment: difference, which was very high on the PEC table, and Entropy: Difference, which was very high on the FOC table, were added to result in a set of five features. The second set simply used the top five features on the FOC table.

A general description of classifiers is provided in the classification section of this chapter. For this study, the neural network used backpropagation with two hidden layers

and sigmoid nonlinearities. The results were compared to a Bayes linear and quadratic classifier, as well as the K-Nearest Neighbor method.

Of the neural network architectures investigated, the one using five input nodes, five nodes in the first hidden layer, two in the second hidden layer, one output node, and the second set of features yielded the best results with a true-positive rate of 73 percent and a false-positive rate of 35 percent. The K-Nearest Neighbor proved superior to the linear and quadratic classifier, yet both fell short of the neural network.

Giger (17:285,289) provides an exhaustive overview of research studies done on calcification detection, as well as mass lesion detection.

2.6.2 Mass Detection. Twenty-six percent of nonpalpable carcinomas appear as masses (17:284). A mass referred for biopsy has a 10 to 20 percent chance of proving malignant (17:283).

Giger et al (58), from the University of Chicago, use a method that identifies deviations from the architectural symmetry of the left and right breasts. Their system makes use of all four conventional mammographic views that are explained in Appendix A. Gray-level thresholding is performed on the individual mammograms before subtraction techniques are used to combine the images. To decrease the number of false-positives, morphological filtering and analysis of size, shape, and distance from border are employed. Radiologists, with the help of this system, have achieved a true-positive rate of 94 percent, with two false-positives per image.

Bischof et al (17:291) detects masses using template matching that first uses median filters to preserve edges while smoothing noise. Using contrast and size features, Bischof has been successful in reducing the number of false-positives per image to 0.2, while maintaining sensitivity at 81 percent.

Brzakovic and colleagues' (17:291) method exercises fuzzy pyramid linking. The scheme applies gray-level thresholding prior to segmentation by multiresolution image processing. Sensitivity is 95 percent, with false-positives reported as some.

Kegelmeyer (17:291) has been developing a system for detecting stellate lesions which uses histograms of edge orientation and Laws texture energy measures. Using test images, this method obtained a sensitivity of 83 percent, with 0.6 false-positives per image. When radiologists used the scheme as a *second-opinion* they increased their sensitivity by an average of 9.7 percent, with an average increase in specificity of 1.5 percent.

2.6.3 Parenchymal Pattern Detection. The structural architecture, and changes to that parenchymal pattern, is another characteristic investigated to detect and classify breast cancer. Mammograms are screened for changes in the layout of breast tissue, or asymmetry between an individual's breasts, that may be associated with breast cancer development. Most methods are based on the textural patterns of the breast and use this to categorize a woman's risk factor. Wolfe (17:292) established four grades of developing breast cancer that are often referred to by other investigators.

Research done by Magnin et al. (17:292) used local texture features, such as entropy, inertia, and local homogeneity. Their results showed that local texture features were not adequate for discriminating between different patterns.

Yaffe et al (17:292) at the University of Toronto, and Caldwell et al (6), have both explored the fractal dimensions of digital mammograms. Results showed that this method can help distinguish between "high-risk" and "low-risk" parenchymal patterns.

Hajnal and colleagues (18) take advantage of different tissue density to sort mammograms. Higher density make reading more difficult and are, therefore, separated for a closer look.

Several studies use similar analysis of asymmetry between the patient's left and right breast to identify features for classification of benign or malignant. These studies classify an abnormality as opposed to the patient's risk.

2.7 Feature Selection

Of the features that are available to a classification system, it is extremely beneficial to identify which provide the most information and those that may be redundant. Some of the features may be correlated, a scaled duplicate, or even useless for classification. By deleting these features, both feature extraction and classification become less computationally demanding. Additionally, limiting the number of features is critical to the amount of training samples required.

Feature saliency, the power a feature has to discriminate, includes traditional methods and new developments. Conventional feature selection includes procedures such as probability of error determination and K-Means clustering.

Using probability of error as a discriminator relies on probability density functions. Comparing the pdf for each class shows which feature is more likely to correctly separate the classes. Features are selected that have less overlap in the class pdfs. Figure 2.5 illustrates that feature 2 is believed to have better discrimination power.

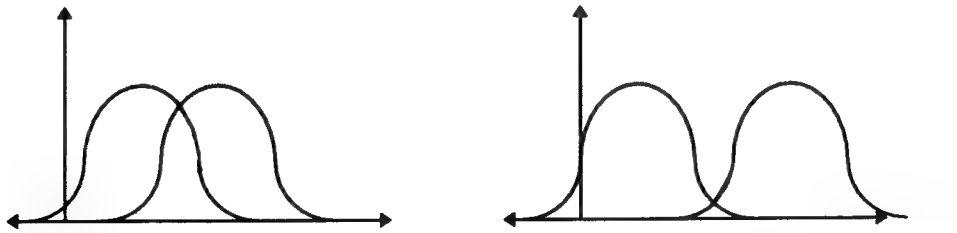


Figure 2.5 Probability of Error Determination

Neural networks can be called upon during this stage in the computer-aided diagnosis system. "The discovery of non-obvious relationships between features may be one the great contributions of neural networks (39:82)." Priddy et al developed a technique for ranking the importance of features for a multilayer perceptron from a Bayesian perspective (37:100). Two example problems compared the resulting saliency measures from their method to the traditional probability of error criterion technique. Similar rankings were noted.

Recent research done by Steppe (50) applies a statistical analysis to develop feature saliency metrics for feedfordward neural networks. Additionally, the Steppe algorithm provides automated architecture and feature selection. Lee and Langrebe (30) base feature reduction on decision boundaries and how those boundaries are oriented to the data.

2.8 Classification Methods

System performance compares true-positive and false-positive classification rates. As shown in Section 2.4.1, the true-positive fraction is the number of correctly identified malignant cases over the total number of malignant. The false-positive fraction divides the number of incorrectly classified benign cases by the total number of benign. The best results will always be a trade-off between these two values.

There exists many of different discriminant functions for use in classification -- those used in this study are explained in the following sections.

2.8.1 Bayesian Classifiers. A Bayesian classifier minimizes probability of error using the method detailed in the Feature Selection section of this chapter. There exist

several classifiers that fall into this category, however, they are conventionally broken down into 2 divisions -- parametric and nonparametric Bayesian classifiers. The parametric classifiers use estimates of the parameters (variables) to form the discriminate function, i.e. mean vector and the covariance matrix, by assuming known *a priori* distributions. Gaussian schemes fall into this category. Nonparametric systems do not assume the underlying class-conditional pattern distributions are known. These schemes include classifiers such as K-Nearest Neighbor and Parzen Window.

2.8.1.1 Gaussian Classifier. Gaussian is the most common classification scheme. It is based on Gaussian distributions centered at the mean of each class. Variance can be calculated per class or class variances averaged to form a grand variance. The covariance matrix may be diagonal, with one variance for each input dimension, or a full matrix.

2.8.1.2 K-Nearest Neighbor. K-Nearest Neighbor (KNN), a Bayesian nonparametric classifier, finds the K closest training samples to the test sample. The test sample is assigned to the dominant class of neighbors. The LNKnet package used in this study uses a Euclidean distance measure and breaks ties randomly.

2.8.2 Artificial Neural Networks. Artificial neural networks represent endeavors to model information processing systems after biological neurons. The only actual connection is the structure of many simple processing elements that operate in parallel. One type of network is referred to as a multilayer perceptron (MLP) (38,50:3).

Several reasons for the use of artificial neural networks in the detection and diagnosis of cancer have been suggested. They learn from experience, work where other algorithms fail, and generalize from training examples to perform well on independent test

data (39:77). Ruck and colleagues (42) dispel several myths and prove that the MLP approximates the Bayes optimal discriminate function. Work was also done by Cybenko (11) that indicates a single hidden layer is sufficient to uniformly approximate any continuous function. Steppe explains Cover's theorem where by examining the degrees of freedom in the model, an upper bound, H, may be set for the number of nodes in the middle layer (50:153).

$$H = \frac{.5P - 1}{M + 1} \tag{2-1}$$

Using the number of training samples, P, and the number of features, M, Equation 2-1 shows how the upper bound is determined. For these reasons, our investigations are performed using one hidden layer with no more than H middle layer nodes.

2.8.3 Post Multiplying the Classifier Output.

During learning, the a priori probabilities, $P(\omega_i)$, for each class are assumed to be proportional to the number of training samples in each class. Hush and Horne (24:22) suggest these probabilities may be adjusted after training. Scaling the a priori probabilities is possible due to the scalar relationship between the a priori probability and the a posteriori probability produced by the classifier, $P(\omega_i / x)$. Equation 2-2 describes the Hush and Horne method to accomplish scaling,

$$\frac{P(\omega_i)_{RW}}{P(\omega_i)_{TRNG}} * P(\omega_i / x) = P_{ADJUSTED} \tag{2-2}$$

where i represents the class, $P(\omega_i)_{RW}$ the a priori class probability in the real world, and $P(\omega_i)_{TRNG}$ the a prior class probability from the training set. The sum of the adjusted probabilities across all classes should be one.

This adjustment could be applied to scale the available data according to individual risk factors or cancer facts such as a suspicious lesion having a 30 percent change of proving malignant.

2.9 Summary

An introduction to breast cancer development, diagnosis, and risk factors has been developed. Considerations important to acquisition technology, database selection, and performance reporting were covered. Segmentation was introduced as a significantly daunting task for breast cancer image processing. Feature extraction methods are cover mass lesions, microcalcifications, parenchymal changes, or any combination of these. Feature selection methods were identified to isolate and remove redundant or useless features. Several classifiers were presented for use in discriminating between malignant and benign cases.

This background information is expanded in the following chapters. Specific methods employed for feature extraction; ASM, *eigenmasses*, and wavelets, are developed in detail in Chapter III. Chapter IV provides database specifications, individual performance for the features extracted, and results of feature selection. Final results for the selected features is summarized in Chapter V.

III. Methodology

3.1 Angular Second Moment

Angular second moment is a feature based on co-occurrence -- a second-order statistical measure of image variation. The co-occurrence matrix provides the basis for a number of textural features (45:273). In the specific problem at hand, the variation being observed is gray-level changes in a mammographic image.

3.1.1 Gray-level Histograms. Figure 3.1 is used as an example image to visualize each definition for a gray-level histogram.

0	1	2	4
3	0	1	2
4	1	4	4
0	1	0	1

Figure 3.1 Gray-Level Values of an Example Image

A first-order gray-level histogram is simply the probability distribution function representing the occurrence of a single gray-level value. The first-order histogram for the example is displayed in Figure 3.2.

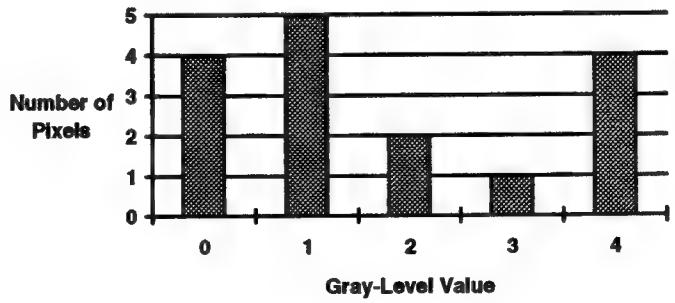


Figure 3.2 First-Order Gray-Level Histogram

Figure 3.3 depicts an actual mammographic image, with Figure 3.4 being the corresponding first-order gray-level histogram.

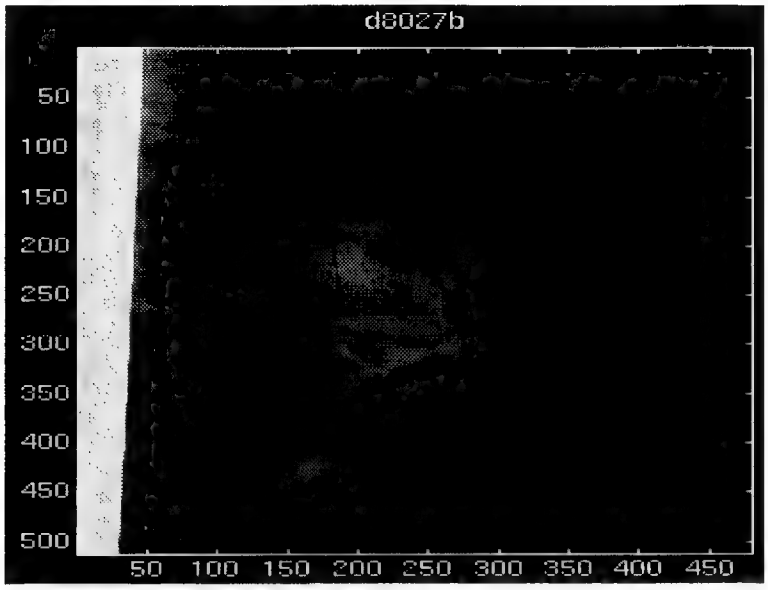


Figure 3.3 Mammographic Image with Microcalcifications

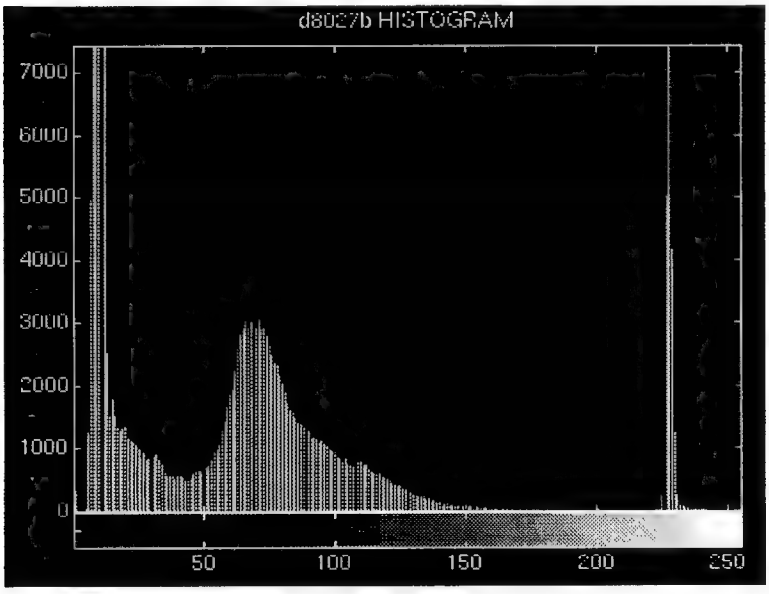


Figure 3.4 Mammographic Image Histogram

The second-order gray-level histogram takes into account the probability of a pair of gray-levels occurring at some given distance vector, $\overline{d}$. Figure 3.5 shows the resulting histogram for various distance vectors. Since an arbitrary distance vector is chosen only horizontal or vertical changes are identified.

d[2,0]	0	1	2	3	4
0	1	0	2	0	0
1	0	1	0	0	2
2	0	0	0	0	0
3	0	1	0	0	0
4	0	0	0	0	1

a) Using a distance vector of [2,0].

d[0,1]	0	1	2	3	4
0	0	1	0	1	0
1	1	1	0	0	1
2	0	1	0	0	1
3	0	0	0	0	1
4	2	1	1	0	0

b) With distance vector of [0,1].

Figure 3.5 Second-Order Gray-Level Histogram

Although displayed as a matrix for ease in viewing, these histograms are actually three-dimensional with the number of occurrence being the out-of-page axis.

3.1.2 ASM Calculation. Angular second moment provides a measure of uniformity (12:1382) and can be calculated from the co-occurrence matrix (the second-order gray-level histogram) by summing the square of all values in the matrix. Shown in equation form:

$$ASM = \sum_{i=0}^{y_q} \sum_{j=0}^{y_r} x_{ij}^{\ 2}$$

For the example, ASM for Figure 3.5a is 12 and for Figure 3.5b is 14. Higher values of angular second moment define higher structural variations (12:1382). Appendix F contains the Matlab code developed to calculate ASM for this study.

3.2 Karhunen-Loeve Transform

The Karhunen-Loeve Transform (KLT), also referred to as principal components, is a mechanism used to determine the directions of variance in a given feature set, as depicted by the eigenvectors in Figure 3.6.

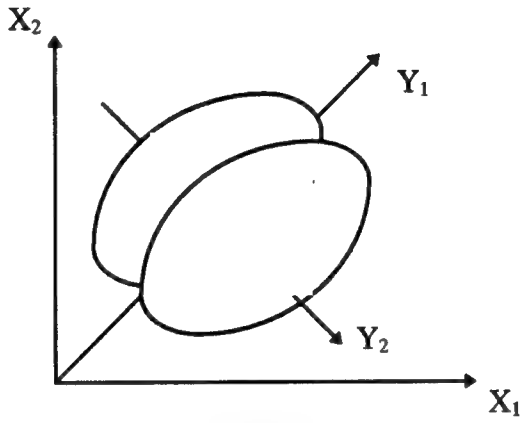


Figure 3.6 Eigenvectors

The covariance matrix summarizes the amount of information in each pixel and is also a second order calculation. By calculating the covariance between one pixel and every other pixel in the image, a determination can be made on the information content of that one pixel. This calculation is done for each pixel location to create the covariance matrix. When a single image is used, the covariance matrix will represent the co-occurrence matrix for all possible distance vectors. Each distance vector diagonally spans the covariance matrix. Since all distance vectors are taken into account, both horizontal and vertical changes are represented.

Based on diagonalization, the KLT is a eigenvalue-eigenvector-based coordinate transformation of the covariance matrix (46:306), with the eigenvectors used as a basis set. By projecting a new vector onto the basis set, named the *eigenmass*, the resulting coefficients can be used as features to classify the data.

3.2.1 Eigenvalues/Eigenvectors. Eigenvalues are the solution for λ in the equation

$$C\bar{x} = \lambda\bar{x} \tag{3-1}$$

where C is a $n \times n$ (square matrix), $\bar{x}$ is a length n column vector, and λ a scalar. By definition, the eigenvector is a scaled multiple of the same vector with no rotation. In the equation above, $C\bar{x}$ is scaled by a factor of λ and has the same direction in n-dimensional space (46:295).

To find the set of non-trivial solutions to Equation 3-1, solve the rewritten form

$$|C - \lambda I| = 0 \tag{3-2}$$

where I is the identity matrix. Since C is a $n \times n$ matrix, there will be n solutions for λ and these are referred to as the eigenvalues. Each eigenvalue has a corresponding eigenvector, $\bar{x}$, that solves Equation 3-1. The eigenvectors will be orthogonal.

3.2.2 Eigenmass Method. Applying the KLT theory to the problem of cancer classification, selected ROIs from the mammograms are used to create the eigenmass. These selected $n \times n$ ROIs are chosen to represent benign and malignant regions. The number of images (ROIs) in the training set is denoted by M. The pixels of each image are placed into column vectors by moving completely across a row before proceeding to the next. The resulting $1 \times n^2$ column vectors, $\bar{x}_i$ (where i is the image number and goes from one to M), will be in the following format.

$$\overline{x_i} = \begin{bmatrix} x_{11} \\ x_{12} \\ \cdot \\ \cdot \\ x_{21} \\ x_{22} \\ \cdot \\ \cdot \\ x_{n^2} \end{bmatrix}$$

The *averagemass* is now calculated by averaging each pixel over all M images.

$$\overline{A}_x = \frac{1}{M}\sum_{i=1}^{M}\overline{x_i} \tag{3-3}$$

$\overline{A}_x$ is a 1 x n^2 matrix and stores the average value for each pixel location. Subtracting the averagemass from each $\overline{x_i}$ yields of set of *centered* image vectors.

$$\sum_{i=1}^{M}(\overline{x_i} - \overline{A_x})$$

This will alleviate the extra component $-\overline{A}_x\,\overline{A}_x{}^T$ from being carried through further calculations. By augmenting all $\overline{x_i}$ vectors in the following manner

$$\mathrm{X} = \left[\overline{x_1}\middle|\overline{x_2}\middle|\overline{x_3}\middle|\cdots\middle|\overline{x_M}\right]$$

an overall n^2 *x M* matrix, X, will be produced to represent all of the training data.

The covariance matrix, *COV*, is computed by

$$COV = \frac{1}{M}\mathrm{X}\mathrm{X}^T = \frac{1}{M}\begin{bmatrix} \sum_{i=1}^{M} x_{i_1}x_{i_1} & \cdots & \sum x_{i_1}x_{i_n} \\ \vdots & \ddots & \vdots \\ \sum x_{i_n}x_{i_1} & \cdots & \sum x_{i_n}x_{i_n} \end{bmatrix}$$

Solving Equation 3-2 for λ yields the set of M eigenvalues. Using each λ in Equation 3-1 will produce the corresponding eigenvectors. To be sure all calculations have been performed correctly, a check is done to ensure the eigenvectors are orthogonal.

Although the Karhunen-Loeve transformation is optimum for signal representation in the sense that it provides the smallest mean square reconstruction error for a given number of features, quite often the features defined by the Karhunen-Loeve transformation are not optimum with regard to class seperability (30:388). For this reason, feature saliency could be done on the eigenvectors to select the ones with the most class discrimination power to represent the KLT basis set.

3.3 Wavelets

Three different wavelets, which are multi-resolution basis transformations that can be used to efficiently and compactly represent different textures, are applied in this study. Several others are suggested in Chapter V as a future area of research. Using computer code written by Myers (34), Daubechies _4, Daubechies_20, and biorthogonal wavelets were applied to the mammographic images. The algorithm takes an image and decomposes it by levels. Each level halves the size of the previous image and creates four new matrices representing the component of the image -- an approximation depiction, and horizontal, vertical, and diagonal detail coefficients. The chosen algorithm only decomposes the approximation depiction further, with the process continuing until single pixel depictions result. Figure 3.7 illustrates this procedure with a simple example, while Figure 3.8 depicts the first level of the decomposition process of an actual mammographic region of interest image.

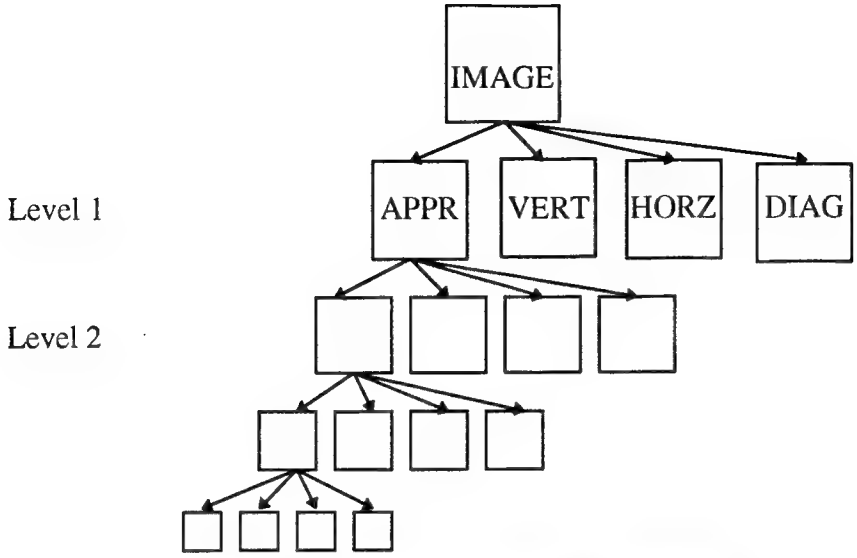


Figure 3.7 Wavelet Decomposition Process

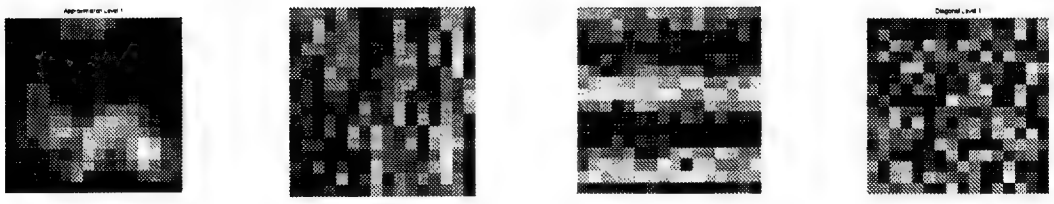


Figure 3.8 Wavelet Decomposition for Mammographic ROI

3.4 Summary

In this chapter, three feature extraction methods have been introduced. Angular second moment has been shown to be a second-order gray-level statistic. Also explored was the novel application of the Karhunen-Loeve Transform to create *eigenmasses.* *W*avelets, having previously only been applied for segmentation, were presented as a feature extraction method. Performance for each of these feature sets is provided in Chapter IV, along with facts on the database applied. Furthermore, feature selections methods are employed to further increase performance on the chosen feature set. Final results are furnished in Chapter V.

IV. Data and Analysis

4.1 Database

4.1.1 Digital. The digital portion of the database was obtained using the LORAD Stereotactic Breast Biopsy system with the cooperation of SouthView Hospital, a training hospital, in Centerville, Ohio. All images were directly acquired as digital (not digitized). Using twelve bits per pixel, this design produces images like those shown in Figure 4.1 and Figure 4.2 that use 4096 gray levels. Image size is either 512 by 512 pixels with resolution of 100 microns, or 1024 by 1024 pixels with 50 microns resolution. X-ray film resolution is 25 microns. Specifics on each file, pathology results for all patients, and LORAD system specifications are contained in Appendix B.

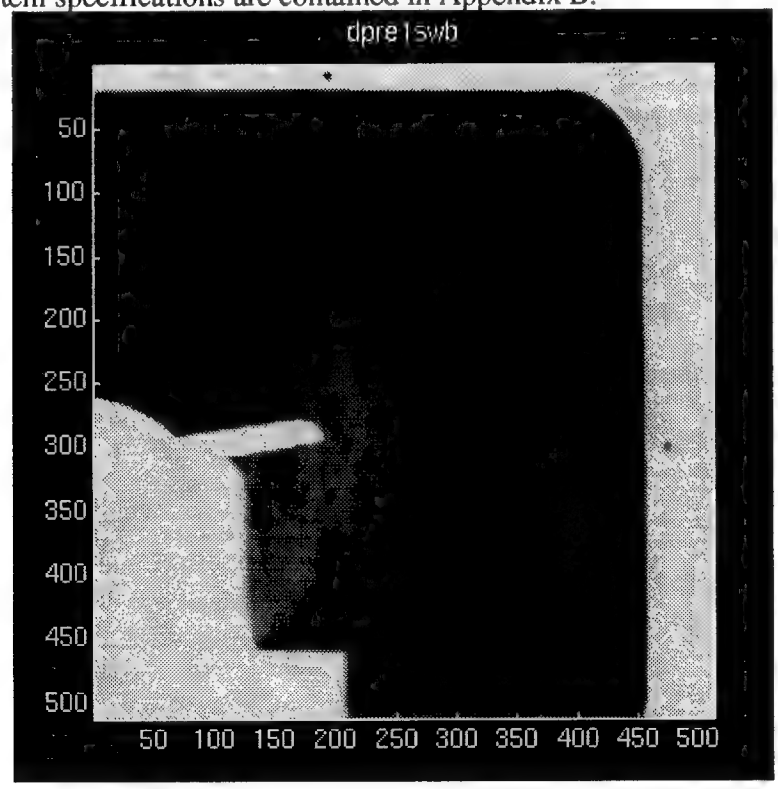


Figure 4.1 Digital Mammogram Containing a Mass Lesion

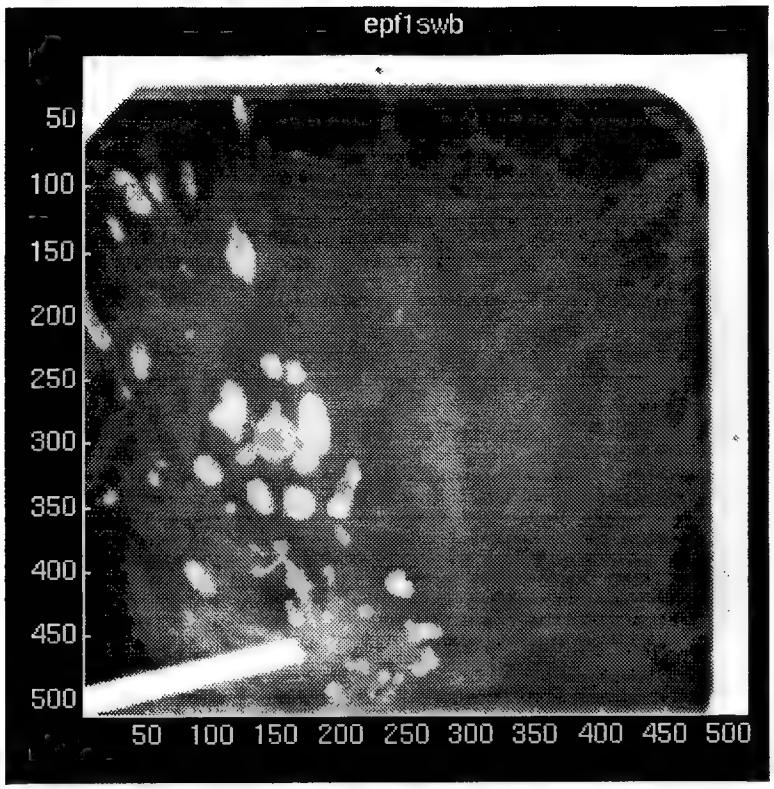


Figure 4.2 Digital Mammogram Containing Microcalcifications

4.1.2 Digitized. Ninety-four "difficult-to-diagnose" mammograms, 63 benign and 31 malignant cases, were obtained from Atam Dhawan, PhD, at the University of Cincinnati. Appendix C contains specifications and points of contact for this database.

These images were digitized at 160 micron resolution. Using 8 bits, the gray-level scale is 0 to 255. In all of these cases, biopsies were performed as a result of identifying microcalcifications. Figure 4.3 portrays one of the resulting images in this database, with the area of concern circled. Image size is 512 by 512 or 512 by 480 pixels.

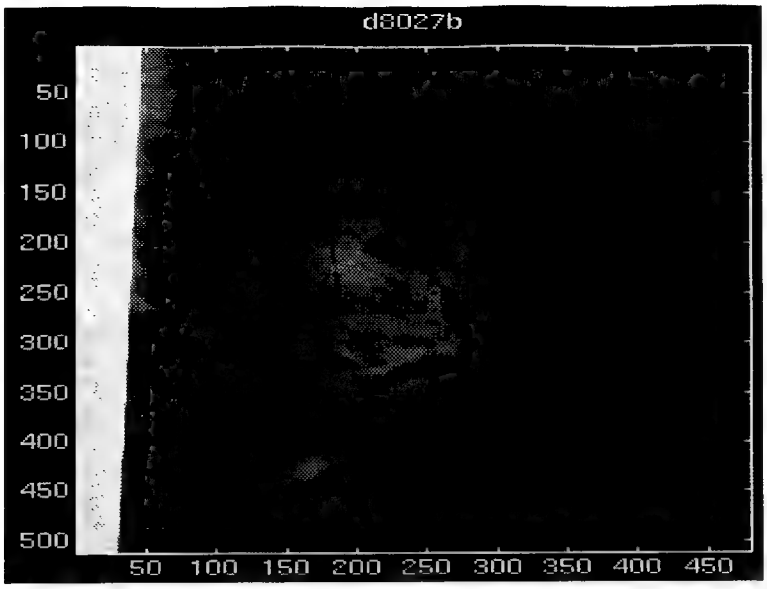


Figure 4.3 Digitized Mammogram with Microcalcifications

The University of Cincinnati group used two binary masks, denoted a and b, to extract sub-regions that contained the area of microcalcifications from each image. The sub-region shown in Figure 4.4 illustrates the resulting image after application of this segmentation technique on Figure 4.3.

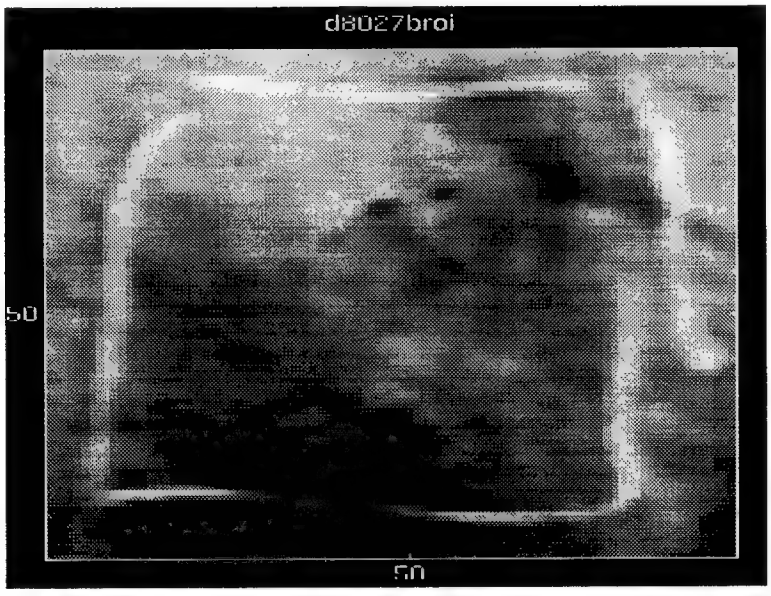


Figure 4.4 Sub-Region Image

The specifics on the method used for this procedure and the digitization of the mammograms is provided in Chitre (9:1380).

4.1.2.1 File Format. The University of Cincinnati images were obtained in raw binary format. Although acceptable for display, these images need to be in ascii format for feature calculations. The Matlab program binary2ascii.m was used to accomplish this task. A copy of this code is provided in Appendix F. Files with the suffix .ascii have been run through this program.

The first four lines of these files contain information specific to each image (done to coincide with the format required for Kelley's (25) Mammogram Image Processing Program). File format is as follows:

line 1 - 98 (for benign) or 108 (malignant)

line 2 - number of rows in the matrix

line 3 - number of columns in the matrix

line 4 - number of ROIs contained in the file

rest of file - matrix containing the gray-level value for each pixel location.

4.1.2.2 ROI Selection. Smaller regions were taken from each sub-region provided by the University of Cincinnati for two reasons: 1) eigenmass and wavelet feature extraction algorithms were developed to handle square matrices, and 2) marks on the original film made by the radiologist had been digitized and were included in the sub-regions. The radiologist had circled the area of concern and this band had also been digitized. In some cases, this band even passed through an abnormality. It was felt that this flaw in the images skewed the raw data. For these reasons, square matrices were selected from the center of each sub-region. The assumption was made that if the

information to determine benign or malignant was in these sub-regions, the same information should also be in any area taken from this sub-image.

Regions 32 by 32 pixels were selected to account for the smallest sub-region in the Cincinnati database and produced images like the one shown in Figure 4.5.

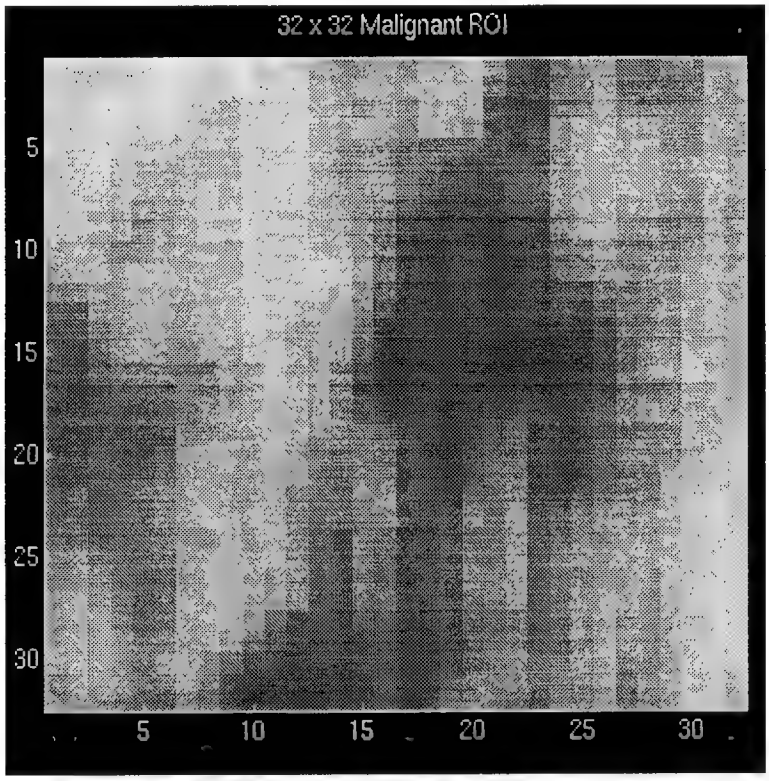


Figure 4.5 Region of Interest Image

The Matlab program sqroi.m, provided in Appendix F, creates the 32 by 32 pixel ROIs and appends the header information required for the Mammogram Image Processing Program. The new file names have roi prefixed to the original file name. These ROIs are the images used throughout this study for feature extraction and classification.

4.1.3 Other Sources. Recent collaboration has also been started with the University of South Florida. Their images are provided courtesy of the H. Lee Moffitt Cancer Center and Research Institute and the Department of Radiology, College of Medicine, University of South Florida (USF). Their 100 images, half containing cancer and half non-suspicious, are in raw format. This database is currently in the AFIT file transfer program directory under pub/for_ckocur/public.

Georgetown University Medical Center is another possible source for a digital database. Doctor Freedman and colleagues are developing full-breast digital mammography. Currently, they can accomplish this for small mammograms (8x10 inches) but not for large mammograms (10x12 inches). Resolution for these images is five line pairs per millimeter or a pixel resolution of 100 microns. In comparison, x-ray film resolution is approximately 20 lp/mm (25 microns). However, size for malignant microcalcifications commonly varies between 100 and 300 microns. Results and the database from their study comparing digital to film mammography will be available after presentation to the Radiological Society of North America (RSNA) in November 1994.

All information currently available on the USF and Georgetown databases, as well as points of contact, are contained in Appendix D.

4.2 Image Display

Both Matlab and Khoros have been used to display and manipulate the mammograms. Although Khoros was originally used to display the images, it takes quite a lot of finesse to load, convert, then display each image. For this reason, the Matlab Mammogram Image Processing Program, developed by Kelley (25), was used for this thesis and is introduced in Section 4.2.2.

4.3 Classification

The Lincoln Labs Neural Network (LNKnet) package was used for classification. It provides a number of algorithms, several of which have been applied to evaluate performance of the CAD system.

The features obtained from each extraction method developed in this study were fed to a variety of LNKnet classifiers. Performance using KNN, Gaussian, and MLP classifiers was explored using a leave-one-out method. This procedure is suggested when the number of samples for a problem is less than 200 (27:35). It takes N samples, trains the classifier on N-1 samples, then uses the remaining one sample to test. Classification is continued in this manner until all N samples have been used as the test sample. Final performance is reported as an average of classification results for all N trails. In LNKnet, this is accomplished using cross-validation with the number of folds equal to the number of samples.

The best possible classification results for each feature set was found by varying certain parameters for each of the classifiers. The number of nearest neighbors was varied in the KNN classifier with K never being greater than the total number of samples from the smallest class. Variance and covariance matrix architecture was manipulated in the Gaussian method. For the MLP, the maximum number of middle nodes was determined using Equation 2-1. Training epoch and the number of middle nodes were varied to observe their effect on performance.

Selected performance for each feature set/classifier combination is displayed in the following section. Each classifier is listed including its architectural makeup. KNN shows the number of nearest neighbors, Gaussian parameters for variance (class or grand) and covariance (diagonal or full) are displayed, and the number of training epochs and middle nodes are shown with MLP.

4.4 Individual Feature Performance

4.4.1 Angular Second Moment. The angular second moment value, following the method outlined in Chapter III, was found for each of the 94 ROIs. This single feature was imported into LNKnet and classification algorithms explored to obtained the results shown in Table 4.1.

Table 4.1 Classification Performance for ASM

	PERCENT CORRECT		
CLASSIFIER	BENIGN	MALIGNANT	OVERALL
1NN	67	52	62
Gauss(C/D)	99	0	66
MLP(100/2)	75	26	59

4.4.2 Eigenmass. Twenty of the 32 x 32 pixel ROIs were used to create an eigenmass basis set. The 40 samples suggested by Suarez (52) was cut in half since reconstruction was not a factor in our problem. Twenty features could then be obtained for each of the remaining 74 ROIs. Table 4.2 shows the results obtained using these features in the LNKnet classifiers.

Table 4.2 Classification Performance for Eigenmass

	PERCENT CORRECT		
CLASSIFIER	BENIGN	MALIGNANT	OVERALL
4NN	93	20	28
Gauss(C/D)	78	34	65
MLP(100/1)	72	39	63

4.4.3 Wavelets. The method described in Chapter III was applied for wavelet decomposition producing 4 results for each level -- an approximation and 3 detail coefficients. The RMS of each was used as a feature to represent the benign and

malignant images in classification. Having started with 32 x 32 pixel ROIs, wavelet decomposition produced 21 features. Feature 1 represented the RMS for the original image, Feature 2 the RMS for the level 1 approximation, Feature 3 the RMS for the level one vertical detail coefficients, Feature 4 the RMS for the level one horizontal coefficients, Feature 5 the RMS for the level one diagonal coefficient, and so forth in the same order down the levels. By keeping track of what each feature represents, the results from feature saliency can also lend some information to which decompositions provide the most information. This is a possible topic for future research discussed in Chapter V.

Using all 94 samples, each distinctive set of wavelet features was classified separately with results displayed in Table 4.3.

Table 4.3 Classifier Performance for Wavelets

DAUBECHIES-4 WAVELET			
	PERCENT CORRECT		
CLASSIFIER	BENIGN	MALIGNANT	OVERALL
4NN	80	13	58
Gauss(C/D)	66	26	53
MLP(300/2)	78	62	73

DAUBECHIES-20 WAVELET			
	PERCENT CORRECT		
CLASSIFIER	BENIGN	MALIGNANT	OVERALL
8NN	86	23	65
Gauss(C/D)	62	39	55
MLP(100/2)	42	58	47

BIORTHOGONAL WAVELET			
	PERCENT CORRECT		
CLASSIFIER	BENIGN	MALIGNANT	OVERALL
8NN	86	23	65
Gauss(C/D)	66	46	59
MLP(200/2)	83	55	74

From this point, however, feature saliency methods can be applied and possible combinations of features tried to increase classifier performance. Final results are disclosed and explained in Chapter 5.

4.5 Feature Selection

The 21 biorthogonal wavelet feature set was chosen to exercise the feature selection techniques. Both saliency rankings and performance with the selected features are investigated.

4.5.1 Probability of Error. The probability of error method for feature selection, introduced in Section 2.7, was applied for a first look at the saliency of each biorthogonal wavelet feature. Both the benign and malignant Gaussian distributions for the same feature were placed on one graph, as illustrated in Figure 4.6.

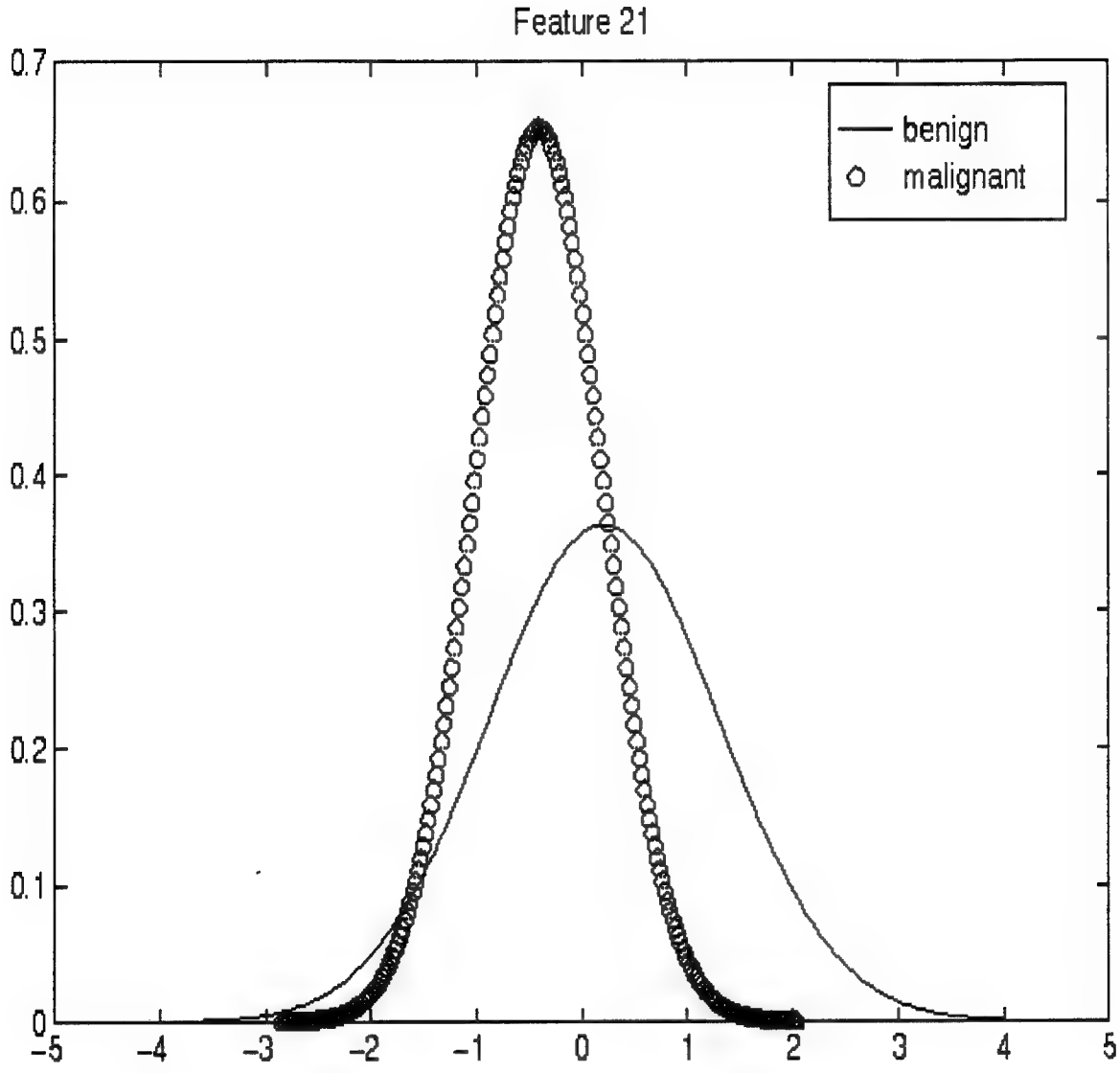


Figure 4.6 Low Probability of Error Feature

The comparison for Features 3, 5, 7, 9, 11, 12, 13, 15, and 21 showed distinguishable benign and malignant distributions, with Features 5 and 21 having the smallest probability of error. However, several features had distributions that overlapped similar to Figure 4.7. Features 1, 2, 6, 10, and 14 where determined to have nearly no saliency (discrimination power).

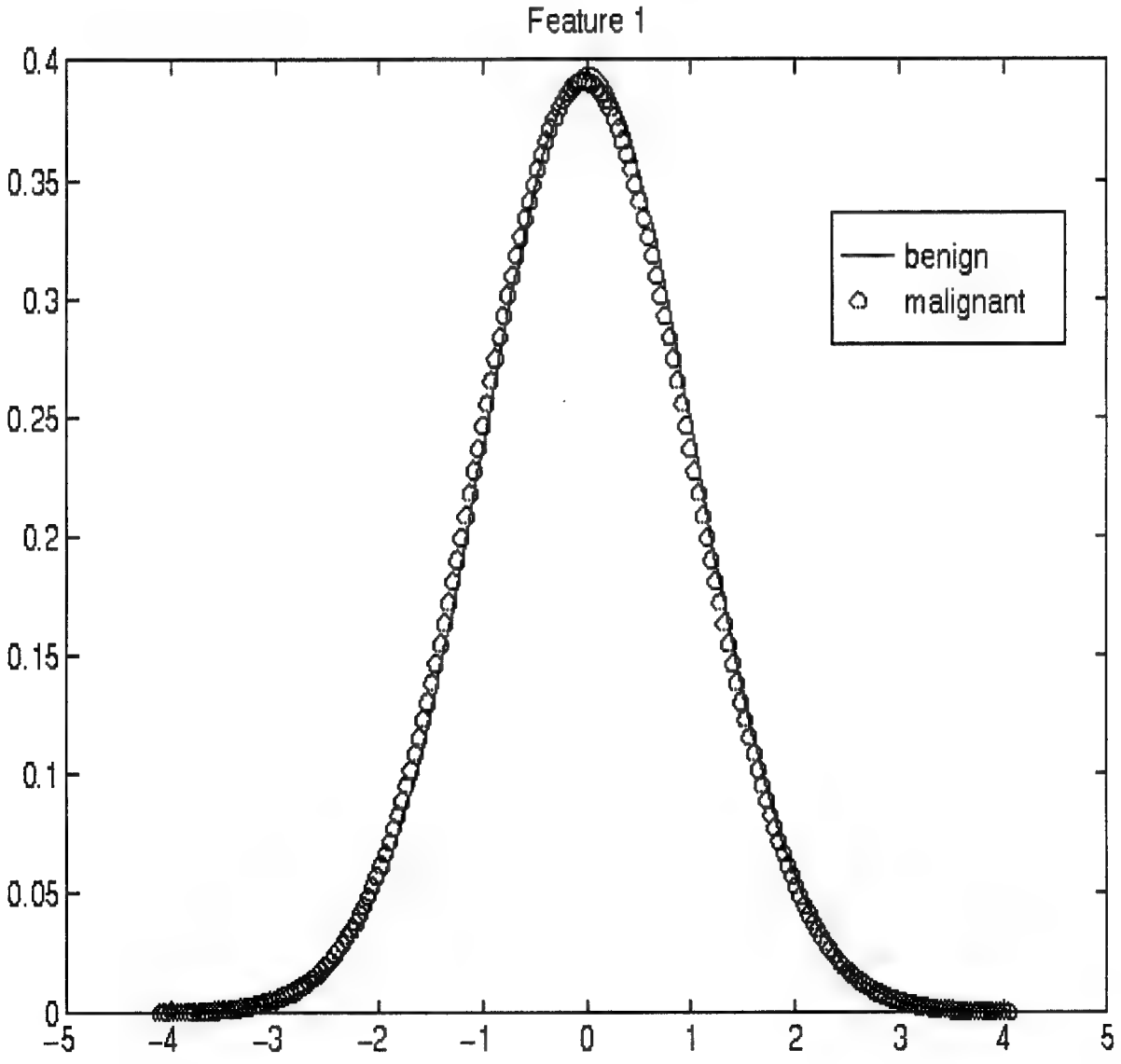


Figure 4.7 High Probability of Error Feature

4.5.2 Steppe Algorithm. Recent research by Steppe (50,51) provides automated model selection for feedforward neural networks. This tool provides feature saliency metrics, feature selection which investigates a good feature subset, and model selection to determine an appropriate network architecture (number of middle nodes). Both derivative-based and weight-based saliency metrics are developed to measure the importance of the feature and provide a final ranking. Feature screening methods identify and eliminate noisy features. By augmenting the original feature set with a known irrelevant, noisy feature statistical analysis can distinguish those features which are of little or no use. It is recommended to eliminate any feature with a mean saliency that falls inside the one-sided confidence interval of mean saliency of the noise (50:107). The Steppe saliency rankings for the biorthogonal wavelet features are shown in Table 4.4.

Table 4.4 Steppe Saliency Means and Ranks

RANK	FEATURE	MEAN
1	21	3.34
2	13	2.016
3	7	1.786
4	19	1.409
5	15	1.307
6	17	0.9071
7	20	0.9017
8	3	0.8522
9	9	0.8245
10	12	0.7662
11	4	0.7076
12	8	0.6862
13	16	0.6797
14	5	0.6617
15	11	0.6459
16	noise	0.3514
17	14	0.247
18	2	0.2306
19	1	0.2145
20	6	0.1942
21	18	0.1813
22	10	0.171

These rankings illustrate that Features 1, 2, 6, 10, 14, and 18 are considered less salient than noise. The Scree Plot in Figure 4.8 was developed using the saliency metric mean that determined the feature rankings.

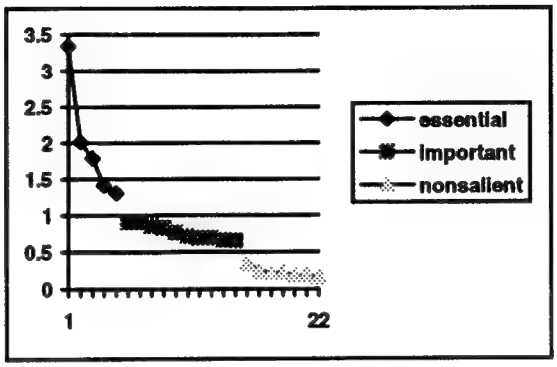


Figure 4.8 Scree Plot of Saliency Means

Using this tool, both essential and nonsalient (useless) features can be identified.

Two algorithms are used in automated model selection. Architecture selection determines a good number of middle nodes, and feature selection statistically investigates a reduced feature set (separate from the prior saliency investigation). The option exists to accept only statistically equivalent reduced models using the likelihood ratio test statistic (50:157). However, allowing the algorithm to reduce the model until one feature remains ensures that reduction did not stop prior to identifying the "best model," which contains a reduced set of features and statistically equivalent classification results to the original model.

The Steppe algorithm was allowed to run on the biorthogonal wavelet feature set until only one feature remained. Table 4.5 provides a synopsis of the trail the algorithm followed in node and feature reduction.

Table 4.5 Steppe Model Selection

# Nodes/Features	Feature Set	% Error	SSE
2n/21f	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21	26	14.34
1n/21f	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21	23.4	15.18
20	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,17,18,19,20,21	25	15.67
19	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,18,19,20,21	14.89	12.66
18	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,19,20,21	14.89	12.42
17	2,3,4,5,6,7,8,9,10,11,12,13,14,15,19,20,21	17.02	13.37
16	2,3,4,5,6,7,8,9,11,12,13,14,15,19,20,21	14.89	13.86
15	2,3,4,5,7,8,9,11,12,13,14,15,19,20,21	13.83	12.62
14	**2,3,5,7,8,9,11,12,13,14,15,19,20,21**	**11.7**	**10.27**
13	2,3,5,7,8,9,11,12,13,14,19,20,21	14.89	13.7
12	2,5,7,8,9,11,12,13,14,19,20,21	19.15	13.91
11	2,5,7,8,9,12,13,14,19,20,21	17.02	13.5
10	5,7,8,9,12,13,14,19,20,21	17.02	13.35
9	5,7,8,12,13,14,19,20,21	20.21	13.61
8	5,7,8,12,13,19,20,21	21.28	14.19
7	5,7,8,12,13,19,21	21.28	14.15
6	7,8,12,13,19,21	20.21	14.66
5	7,8,13,19,21	22.34	14.45
4	7,13,19,21	22.34	14.97
3	13,19,21	25.53	15.11
2	13,21	26.6	15.07
1	21	34.04	19.04

The graphical depiction seen in Figure 4.9 displays how the classification percent error changed as the model was reduced. The architecture of one middle node and 14 features produced the best classification.

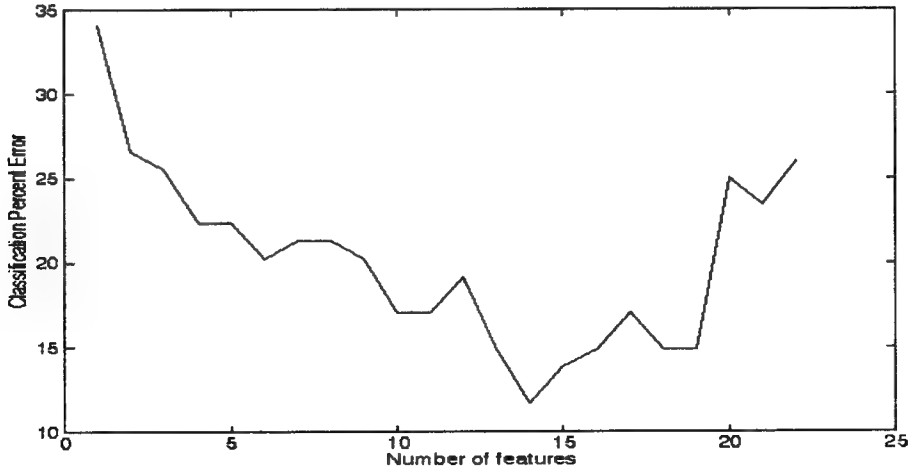


Figure 4.9 Percent Error Change During Model Reduction

4.5.3 Lee and Langrebe Theory. The Lee and Langrebe (30) method for feature selection is based on the decision boundaries and how they are oriented to the data. The process retains informative features while eliminating redundant features. Their objective is to predict the minimum number of features that can achieve close to the same classification accuracy as in the original space (30:388).

To apply this theory, a Gaussian classifier (diagonal covariance/per class variance) was used to identify misclassified samples when testing on all 94 training samples. The mean and variance of each class was calculated for all 21 normalized biorthogonal wavelet features. A new data set with only the correctly classified samples was created. (Note, that the misclassified samples are already ignored. The intention is to attempt to get the correct ones correct with a reduced feature set). These parameters were fed to the algorithm developed by Eisenbies (16) to implement the Lee and Langrebe theory.

Although not the final objective on Lee and Langrebe, the resulting eigenvalues and eigenvectors alone gave us insight into the relative importance of several features. Each eigenvector, which represents a combination of all 21 features, was examined to identify which features were most dominate by examining the magnitude of each feature.

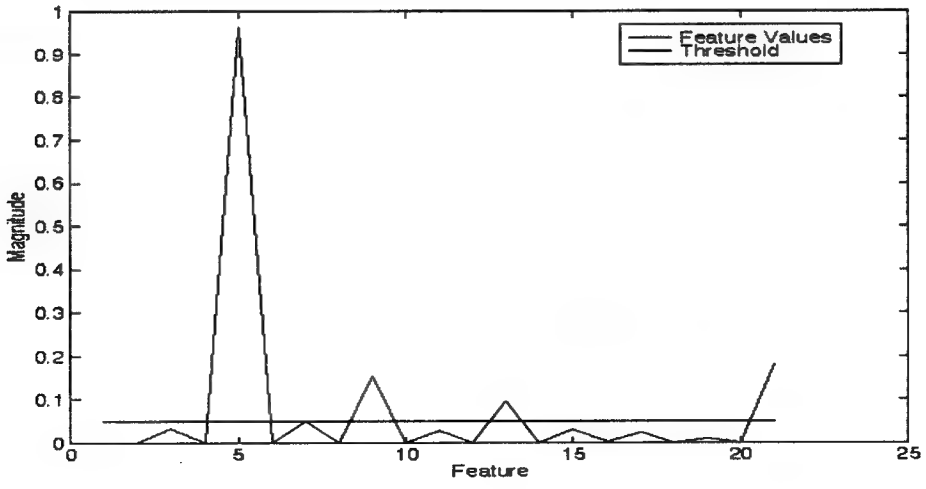


Figure 4.10 Feature Combination for the Largest Eigenvalue Vector

From Figure 4.10 it can be seen that the largest influence in this vector comes from Features 5, 21, 9, 13, and 7 (in order of reducing magnitude). Features 1, 2, 4, 6, 8, 10, 12, 14, 16, and 18 show the small impact. Figure 4.11 lists the most significant features for the dominate eigenvectors.

Dominate Features for Maximum Eigenvectors				
5	**9**	3	**21**	15
21	13	**21**	3	**21**
9	**21**	**9**	13	13
13	7	12	7	**9**
7	5	15	15	
	17	17	**9**	
	11	7	5	
	3		20	
	15			

Figure 4.11 Lee and Langrebe Eigenvector

It is plain to see that Feature 9 and 21 are significant magnitude features common to most eigenvectors. Features 1, 2, 6, 10, 14, 18, and 20 are commonly considered insignificant. This saliency measure for the features is consistent with the finding using the Steppe method.

The eigenvectors represented the basis set. A transformation was done by multiplying the samples (the original data file) by the basis set. In this new space, classification performance could be investigated with the data projected onto a new axis set. As Figure 4.12 shows, classification accuracy can reach nearly 90 percent (of the previously 100 percent correctly classified samples) with only the first three eigenvectors.

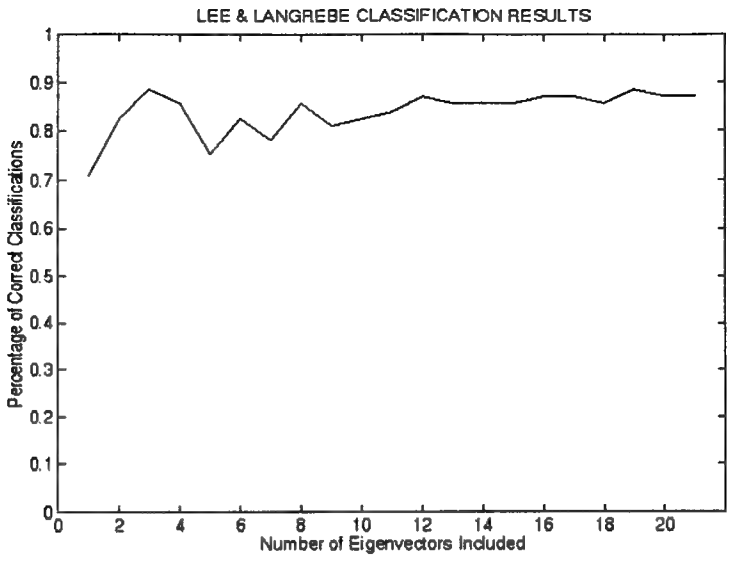


Figure 4.12 Lee & Langrebe Classification Results

4.6 Summary

Database specifications and concerns have been reported for both the digital and digitized databases. ROI selection process was identified and performance using the ROIs with each separate feature set is provided. Feature selection methods were employed on the biorthogonal wavelet set in reduce feature dimensionality and possibly increase system performance. Is was shown, and is summarized in figure, that both the Steppe algorithm and the Lee and Langrebe method produced consistent results in identifying both dominant and useless features. Chapter V highlights the benefit from applying feature selection and reports the final performance of the system.

Table 4.6 Feature Selection Summary

	Dominant Features	Irrelevent Features
Probability of Error	5,21	1,2,6,10,14
Steppe Saliency	21,13,7,19,15	1,2,6,10,14,18
Lee & Langrebe	9,21,7,13,15	1,2,6,10,14,18

V. Results and Conclusions

5.1 Summary

This research began with an introduction to basic cancer facts and diagnostic traits. A review of existing pattern recognition and image processing techniques common to breast cancer research was presented. With this foundation, development of angular second moment as a second-order gray-level histogram feature provided baseline accuracy for a known feature. Based on the probability of a pair of gray-levels occurring at a given distance vector, ASM was calculated by summing the square of each value in the co-occurrence matrix, as shown in Section 3.1.2. The best results achieved from the ASM feature was an overall 62 percent correct classification rate.

In an effort to improve upon the classification performance, two novel feature extraction methods, *eigenmass* and wavelets, were introduced to the field of computer-aided breast cancer diagnosis. Application of the Karhunen-Loeve Transformation, which obtained the *eigenmass* as shown in Section 3.2.2, includes additional information that is lost in the simple second-order statistics. ASM calculations are done on one chosen distant vector. *Eigenmass*, however, retains information on all distant vectors. By projecting a new vector onto the *eigenmass* basis set, the resulting coefficients can be used as features to classify the data. *Eigenmass* features achieved an overall 65 percent correct classification rate.

Wavelet features were constructed by decomposing each image by levels. Each level halved the size of the previous image and created four new matrices -- an approximation signal, and horizontal, vertical, and diagonal detail coefficients. Decomposition continued only on the approximation until single pixel depictions resulted.

The root mean square of each of these was used as a feature resulting in a set of 21. Daubechies_4, Daubechies_20, and biorthogonal wavelets were all investigated with the best overall performance of 74 percent correct classifications being obtained by the biorthogonal wavelet.

Feature selection methods were examined to improve the classification results of the biorthogonal feature set. Statistical analysis provided an introductory look at discrimination power. The Steppe algorithm provided a saliency ranking for each feature and automated model selection obtained the optimum number of middle nodes as one and the feature set to contain 14 of the original 21 features. Lee and Langrebe theory was introduced as a feature selection technique based on decision boundaries. These methods were compared and isolated useless features for removal.

The biorthogonal wavelet features provided an overall 74 percent correct classification in the distinction between benign and malignant lesions found in mammograms. The application of feature selection techniques increased system performance. Employment of the Steppe algorithm for automated feature selection and neural network architecture resulted in increasing CAD system performance to an overall 88 percent correct classification rate. The final model showed that in fourteen dimensional space the biorthogonal wavelet data was linearly separable to 88 percent accuracy.

5.2 Conclusions

Advances and contributions are made with the introduction of two novel feature extraction methods that have never before been seen in cancer diagnosis. Additionally, feature selection techniques have been investigated, compared, and validated; transforming adequate discrimination power into promising classification results.

Eigenmass features have provided consistent classification performance of 65 percent with a limited database. Wavelets, which prior to this study had only been exploited for their segmentation and preprocessing benefits, are introduced and shown to be a robust feature extraction technique, providing a correct classification rate of 74 percent on difficult-to-diagnose cases.

The procedures introduced in this study provide new insight to computer-aided diagnosis. Having performed on a database limited in size and containing imperfections, classification performance could be improved, and increased confidence in the diagnosis ensured, by applying these methods to a sound database.

The contribution of feature selection is seen in the stark improvement from 74 to 88 percent in the discrimination power provided by the reduced feature set. The Steppe algorithm and Lee and Langrebe technique showed consistent identification of dominant and useless features.

5.3 Recommended Future Research

A suggested topic for future research is to exploit the multiple views already in use for localization in stereotactic needle-biopsy. This technology was introduced in Section 2.3.3.1. The information contained in additional views of the same area of concern can no doubt provide more information to a computer-aided diagnostic system. Additionally, three-dimensional imaging of the breast is in its infancy. This, too, is a must for future research in CAD.

Segmentation continues to be a difficult portion of computer-aided diagnosis. Manual segmentation was performed for this endeavor but it is the hope that this system will be refined to included automated detection of the abnormality. The ultimate objective in this area is for the CAD system to identify possible areas of interest *to* the radiologist.

Described earlier in this study, risk factor analysis could add additional information to a CAD system. By varying the classification thresholds, a system could be developed that may more accurately represent women on an individual basis.

Wavelet application and analysis was introduced in this project. However, having discussed the topic with leading wavelet experts at the Air Force Institute of Technology, it is suggested that avenues continue to be explored using wavelets designed for discrete, finite sized images. Wavelets, such as malvar (designed at AFIT) and cubic spline, work around edge problems. Additionally, wavelet packets which decompose each sub-image, instead of just the approximation, may provide added information.

A separate investigation into which wavelet features provided the most discrimination power may render insight into the type of coefficients (diagonal, horizontal, etc.) needed to distinguish between benign and malignant lesions.

Collection of a variety of databases continues to be an ongoing endeavor. It is suggested, however, that special consideration be given to obtaining images containing mass lesions. The digitized database now being used is limited to microcalcifications. For this reason, the features demonstrated in this study have not been investigated for possible contributions to the classification of mass lesions.

APPENDIX A. TERMS

Benign: not cancerous.

Carcinoma: cancer; malignant lesion.

Circumscribe: "to confine within bounds; restrict (12:275)."

> **Circumscribed Lesion:** a mass lesion with sharply demarcated or well-defined borders in contrast to a mass with indistinguishable borders where it is difficult to tell where the mass starts or stops (can be called a poorly circumscribed lesion).

Contrast: "to show differences when compared (12:318)."
Sanders (24:98) defines contrast as how well a target stands out from its background. Luminous contrast = (Lmax-Lmin)/Lmax.

Conventional Mammographic Views (4): right and left craniocaudal (head-to-tail); right and left mediolaterial oblique (side-to-side).

In Situ: "in place"; very early stages. This refers to cancerous cells that are where they where originally located; early stage of cancer that has not become invasive.

Invasive: "Tending to spread, esp. tending to invade healthy tissue (12:674)."
A type of cancer that is no longer in situ, but has spread beyond the location of the original normal cells. Usually means that the cancer has spread to the blood or lymphatic vessels and therefore can be spread to anywhere in the body.

Lesion: "General term for any visible, local abnormality of tissue (e.g. an injury, wound, boil, sore, rash) (11:256)."

Malignant: cancerous.

Mass: A lesion which is a collection of contiguous cancerous cells in a local area which when x-rayed has increased density over a relatively large area (approximately 1 cm or greater).

Menarche: start of menstrual periods.

Microcalcification: A lesion which shows up on the mammogram as a small (1 mm range) area of increased density and corresponds to calcium in the tissue.

Nonpalpable: "Not perceivable by touch (11:333)." A mass that cannot be detected through physical exam.

Nulliparity: never having borne children.

Parenchyma: "Functional part of an organ, apart from supporting or connective tissue (11:337)." Normal tissue for that organ (normal breast tissue) that is functional.

Parenchymal Pattern: Layout or design of that tissue.

Results:

Positive: a malignant outcome.
Negative: NOT having cancer.

False-Negative: A benign diagnosis when the lesion is actually malignant.
False-Positive: A malignant diagnosis when the lesion is actually benign.
True-Negative: A benign diagnosis when actually benign.
True-Positive: A malignant diagnosis when actually malignant.

Sensitivity: Ability to correctly identify cancer; true-positive. A sensitivity reported as 90% indicates 9 out of 10 cancers correctly identified.

Specificity: False-positives. A specificity reported as 60% represents 9 out of 15 benign cases correctly recognized.

Stellate: "Arranged or shaped like a star; radiating from a center (12:1193)."

Stellate Lesion: Spiculated mass. A mass that appears to look like a star.

APPENDIX B. DIGITAL STEREOTAXIC MAMMOGRAPHIC IMAGES

Header: None
Format: Binary
Size (vertical dimension x horizontal dim): 512 x 512 or 1024 x 1024
Bits per pixel: 12 per 16
grey levels: 4096
Resolution: 50 microns

FILENAME	SIZE	BIOPSY RESULTS
PATIENT A:		
an1swb	1024 x 1024	Malignant mass.
an2swb	(All files	
apostswb	from this	
apre1swb	patient).	
apre2swb		
as2swb		
as3swb		
asswb		
ast1swb		
ast2swb		
PATIENT B:		
bpf11swb	1024 x 1024	Benign calcifications
bpf12swb	(All files	
bpf21swb	from this	
bpf22swb	patient).	
bpost3swb		
bpostswb		
bpre1swb		
bpre2swb		
bpst1swb		
bpst2swb		
bs2swb		
bsswb		
bst1swb		
bst2swb		
PATIENT C:		
		Melanoma
PATIENT D:		
dn1swb	512 x 512	Malignant Mass
dn2swb	512 x 512	
dpostswb	1024 x 1024	
dpre1swb	512 x 512	
dpre2swb	512 x 512	
dsswb	1024 x 1024	
dst1swb	512 x 512	
dst2swb	512 x 512	

PATIENT E:

epf1swb	512 x 512	Benign microcalcifications
epf2swb	512 x 512	
epostswb	1024 x 1024	
epre1swb	512 x 512	
epre2swb	512 x 512	
esswb	1204 x 1024	
est1swb	512 x 512	
est2swb	512 x 512	

FILENAME MEANING

first letter: patient ID

n = needle_placement
post = post
pre = prefire
s = scout
st = stereo_scout
pf = postfire
pst = poststereo

swb: filename with swb have the bytes swapped to the original form
(in other words, bytes were swapped for data storage and have been "unswapped" for processing).

PATIENT E					
ROI NAME	IMAGE	SIZE	MAL/BEN	CANCER TYPE (COMMENTS)	ASM
.roi1	est2swb	512	ben	probable tissue; between calcs	180
roi2				prob. calc.	182
roi3				possible tissue (far from calcs)	184
roi4				poss calc (small round; away from where biopsied)	182
roi5				poss. large calc - homogeneous	182
roi6				poss large calc - heterogeneous	198
roi1	epre2swb	512	ben	probable tissue between calcs	178
roi2				probable calc	176
roi3				tissue far away	178
roi4				IGNORE -- bad data	
roi5				small possible calc (far away)	180
				* no probables for epf2swb -- needle in way	
roi1	epf2swb	512	ben	close possible tissue	176
roi2				close possible calc	180
roi3				far possible tissue	178
roi4				far possible calc	180
roi1	est1swb	512	ben	probable calc	176
roi2				probable tissue	190
roi3				IGNORE --bad data	
roi4				far possible tissue	182
roi5				far possible calc	178
				* no probables for epf1swb -- needle in way	
roi1	epf1swb	512	ben	close possible tissue	184
roi2				close possible calc	188
roi3				far possible tissue	182
roi4				far possible calc	180
roi1	epre1swb	512	ben	probable tissue	184
roi2				probable calc	182
roi3				possible calc	192
roi4				far away tissue	182
				probable = site of biopsy	
				possible = not as sure	

PATIENT A

J.T. FARRELL, D.O. H.D.STAHL, D.O.
V.H. GREGORY, D.O. J.C. WEISS, D.O.

MAMMOGRAPHY:

Bilateral examination was performed with CC, MLO and ML views. Additional CC and ML MAG compression views were done in the right upper/outer quadrant. A previous study from 6/8/88 was reviewed along with an exam from 6/8/88 and another study from 12/28/88.

Since the prior studies, fatty involution is noted. Bilaterally, there is decreasing fibroglandular tissue. There is an asymmetric density in the upper/outer quadrant of the right breast which measures approximately 12x9 mm. in size. In the axillary oblique projection, there was some questionable architectural distortion associated with this. This was less clearly seen in the region of the abnormal density in the other projections (CC and ML views). Also, it is less clearly apparent in the compression magnification views that were done in the CC and ML projection. This density was present retrospectively on viewing the old films from 1988. The changes suggesting architectural distortion were not clearly apparent then, but the lesion has definitely not increased in size. By itself, without the prior studies, I would grade this lesion as suspicious. The fact that it has changed little over a five-year interval would tend to speak toward benign disease. If however anything is palpable in the upper/outer aspect of the breast at approximately the 10:00 to 11:00 position, it should be removed. This lesion should be amendable to stereotactic biopsy.

There are dermal calcifications identified inferiorly and medially in both breasts. Some of these clearly project into the area of the skin surface. Ductal prominence is seen in the subareolar area bilaterally. This is a symmetrical finding. No dominant masses are seen on the left breast. No abnormalities of the axillary folds are seen.

SUMMARY:

1. THERE IS A 1 CM. IRREGULAR DENSITY IN THE UPPER/OUTER ASPECT OF THE RIGHT BREAST WHICH HAS CHANGED LITTLE OVER A FIVE-YEAR INTERVAL, BUT DEMONSTRATES SOME MAMMOGRAPHIC CHARACTERISTICS WHICH ARE CONSIDERED SUSPICIOUS. FOR THAT REASON, I WOULD RECOMMEND EITHER SURGICAL EXCISION FOLLOWING NEEDLE LOCALIZATION OR STEREOTACTIC BIOPSY. IN PARTICULAR, IF

CONTINUED

JAN 29 1994

RADIOLOGIST [redacted]

X-RAY COPY

PATIENT A

(A) Mass + Cath's

#good up cancer (mass)

Biopsy No. S94-432 2/16/ 1994

-GROSS EXAMINATION

Submitted in formalin and labeled right breast upper outer quadrant mass are hemorrhagic pieces of tissue which range from 2 mm. to 1.2 cm. in greatest dimension. Each piece is approximately 1 mm. thick. Submitted in toto.

MICROSCOPIC EXAMINATION

Sections of the female breast reveal an infiltrating grade II ductal carcinoma. The tumor infiltrates in tubules and cords into the fat. The tumor is composed of large cells with vesicular nuclei showing prominent nucleoli and pink cytoplasm. It has a fibrous stroma. Small foci of hemorrhage and calcification are seen in the tumor.

PATHOLOGIC DIAGNOSIS 9

Infiltrating, grade II ductal carcinoma of the right breast 0402-8503

PATIENT B

STEREOTACTIC BREAST BIOPSY:

Two films were submitted. There is documentation of stereotactic biopsy o microcalcifications in the upper outer quadrant of the right breast. Follo up specimen radiography confirms the presence of calcification with th specimen.

IMPRESSION:

DOCUMENTATION OF STEREOTACTIC BIOPSY. RIGHT BREAST.

J. BENSELER, D.O /ah
dict. 2/19/94
trans. 2/19/94

FEB 19 1994

RADIOLOGIST [redacted]

X-RAY COPY

PATIENT B

benign

Biopsy No. S94-451 2/18/ 1994

-GROSS EXAMINATION

A number of fragments of red-tan tissue are submitted in toto. The largest fragment measures 0.8 cm. in length.

*MICROSCOPIC EXAMINATION

Sections of breast tissue reveal fibrocystic changes. Some ducts are lined by apocrine type epithelium. There is focal hemorrhage. One area exhibits a small fragment of skin and subcutaneous tissue. Calcifications are not seen.

PATHOLOGIC DIAGNOSIS

Fibrocystic changes, stereotactic needle biopsy of upper outer quadrant of right breast 4

benign

PATIENT C

(C) Vas

melanoma

Biopsy No. S93-2150 8-4- 1993

GROSS EXAMINATION

A number of yellow-tan fragments are submitted in toto for microscopic analysis.

MICROSCOPIC EXAMINATION

Sections of breast tissue reveal metastatic, heavily pigmented melanoma. There is marked necrosis, focal hemorrhage and infiltration by numerous chronic inflammatory round cells. In some areas mitotic activity is quite evident.

Previous biopsy Southview Hospital, S91-2477, revealed heavily pigmented superficial spreading malignant melanoma with polypoid configuration (early level IV invasion, 1.52 mm. thickness), skin of mid-thoracic paraspinal area left.

PATHOLOGIC DIAGNOSIS

Metastatic, heavily pigmented melanoma, stereotactic needle biopsy of right breast at 1 o'clock position 9 0402-8723

PATIENT D

D

*good rep. cancer ^ (mass)

H.D. STAHL, D.O.
E.A. KOPPLE, D.O. J.C. WEISS, D.O.

MAMMOGRAM:

Bilateral low dose film mammography was performed in craniocaudal, oblique and true lateral projections with orthogonal compression magnification views on the left.

Both breasts contain fatty tissue. No mass or microcalcifications of malignant nature were seen on the right.

On the left, there is approximate mass at the approximate 2 o'clock position. With the spiculation, the lesion measures in the range of 1 cm. maximal diameter with the mass itself potentially smaller. No associated microcalcifications were confirmed.

Breast contours were not remarkable.

IMPRESSION: LEFT BREAST MASS SUSPICIOUS FOR CARCINOMA. APPROPRIATE DEFINITIVE INVESTIGATION IS IN ORDER.

COMMENT: Case discussed with Dr. Hathaway on 7-30-93. Considerations for follow-up include stereotactic needle biopsy and excisional biopsy.

JCW/kg
d 07/30/93
t 07/31/93

AUG 2 — 1993

RADIOLOGIST _____

X-RAY COPY

[redacted]

PATIENT D

Biopsy No. S93-2192 8-6- 1993

GROSS EXAMINATION

A number of Yellow-tan fragments of tissue are submitted in toto for microscopic analysis. The largest core measures 1.5 cm. in length.

Frank Brecher, D.O.

MICROSCOPIC EXAMINATION

Sections reveal cores of breast tissue presenting infiltrating duct carcinoma. There is an accompanying desmoplastic reaction. The tumor infiltrates fatty tissue. In addition, a fragment of skin is present.

PATHOLOGIC DIAGNOSIS

Infiltrating duct carcinoma, stereotactic needle biopsy of upper outer quadrant of left breast 9 0403-8503

Received: 8-6-93
Completed: 8-9-93
GH-109 REV. 10-81

FB/wp

Frank Brecher D.C

Frank Brecher, D.O. , Pathologist

PATIENT E

E.A. KOPPLE, D.O. J.C. WEISS, D.O.

BILATERAL BREAST MAMMOGRAM:

The patient indicates she has average caffeine usage and is post-menopausal as of 1954. A family history of breast cancer includes 1 daughter in her late 30's. The patient does currently take Premarin. The patient indicates some discomfort in the left breast with no palpable masses. Bilateral nipple retraction was noted on visual examination.

Low dose film-screen mammography was performed in craniocaudal and axillary projections. This was the initial base line examination for this patient and no prior films are available for comparison.

There are heavy dense dystrophic calcifications projecting in the retroareolar areas of the right and to a greater degree left breast. These calcifications have a ductal type orientation. They are large and greater than .5 mm in size for the most part which suggests a benign circumstance. However, these are not solid, with central lucent areas present which is an atypical appearance for benign secretory calcifications and becomes a more suspicious finding. Additionally some of the smaller calcifications in the retroareolar area show some irregularity. More spherical macrocalcifications are also present and contained in this area which may be indicative of galactoceles or fat necrosis. There is some cicatrization and retraction in the retroareolar area, with inversion of the nipple. No other densities within the breasts are identified. The calcifications in the left breast are predominantly involving the inferior medial quadrant.

The right breast also suggests some nipple retraction with multiple small punctate calcifications in the immediate post-areolar region. These calcifications are far fewer in number than the comparison left breast. Some macrocalcifications are noted as well as smaller punctate microcalcifications. No other mass densities are seen. The breasts maintain a primarily fatty involutional density pattern.

CONTINUED:

AUG 23 1993

B-11

RADIOLOGIST

X-RAY COPY

PATIENT E

IMPRESSION:

1. BILATERAL RETROAREOLAR CALCIFICATIONS ARE PRESENT, MUCH MORE NUMEROUS I THE LEFT BREAST EXTENDING INTO THE INFERIOR MEDIAL QUADRANT OF TI BREAST. SOME OF THE APPEARANCE OF THESE CALCIFICATIONS MITIGATE AGAINS BENIGN SECRETORY ETIOLOGY, ALTHOUGH THEY DO FOLLOW A DUCTAL ORIENTATIC PATTERN.
2. BECAUSE OF THE ASSOCIATED FINDINGS OF ARCHITECTURAL DISTORTION AI NIPPLE RETRACTION, STEREOTACTIC BIOPSY OR SURGICAL BIOPSY IS RECOMMENDE IN THE LEFT BREAST. IF THIS YIELDS A POSITIVE ETIOLOGY FOR CANCER, TI RIGHT BREAST RETROAREOLAR AREA SHOULD ALSO THEN BE BIOPSIED.

W.M. MEYERS, D.O./cmw
DICT.: 08/21/93
TRANS.: 08/22/93

RADIOLOGIST _____

X-RAY COPY

[redacted]

PATIENT E

benign calc.

Biopsy No. S93-2651 9-17- 1993

GROSS EXAMINATION

A number of gray-tan cores of tissue are submitted in toto. The largest core of tissue measures 0.5 cm. in greatest extent.

Frank Brecher, D.O.

*MICROSCOPIC EXAMINATION

Sections reveal a few fragments of fibrofatty tissue presenting focal hemorrhage. Mammary ducts and mammary lobules are not seen.

PATHOLOGIC DIAGNOSIS

Fibrofatty tissue, stereotactic needle biopsy of left breast 1

Received: 9/17/93
Completed: 9/20/93 FB;jr
GH-109 REV. 10-81

Frank Brecher, D.O., Pathologist

APPENDIX B. LORAD System Specifications

Digital Image Receptor:
- size: 14" x 10" x 6'
- active image area: 50 mm x 50 mm
- image device: 1024 x 1024 picture element (pixel) CCD
  (configurable as 512 x 512 array for increased sensitivity)
- maximum resolution: 20 pixels per millimeter in X and Y axis
- exposure range: 0.1 to 5.0 seconds
- synchronization: interlocked with x-ray control signal

DSM Computer:
- microprocessor: 80486 @ 50 MHz
- available RAM: 16 MB
- drive controller: 32 bit
- fixed disk drive: 240 MB
- floppy disk drive: 1.44 MB, 3.5"
- optical disk drive: 600 MB, 5.25"
  (Northgate Computer Systems, Inc.
  SONY Magneto Optical Disk EDM-ADAIS
  512 byte/sector rewriteable)

DSM Monitor:
- monitor: 21" high resolution monochrome
- display area: 16" max horizontal, 12" max vertical
- horizontal scan rate: 48 to 108 kHz (auto adjust)
- vertical scan rate: 60 to 80 Hz (auto adjust)
- video bandwidth: 200 MHz
- horizontal resolution: 1024 to 2048
- vertical resolution: 768 to 1536

System Features:
- 5 cm x 5 cm tissue image area
- operator control of display contrast and luminance window
- choice of 512 x 512 or 1024 x 1024 bit resolution display
- menu control of acquisition, display, storage retrieval, and processing functions
- near real time display
  (3 sec from exposure to display in 512 mode)

APPENDIX C. Digitized Database

Points of Contact:

Dr. Atam Dhawan
University of Cincinnati
Department of Electrical Engineering and Computer Engineering
(513)556-6297
electronic mail: adhawan@ece.us.edu

Yateen Chitre, PhD Student
University of Cincinnati
Department of Electrical Engineering and Computer Engineering
(513)556-6297
electronic mail: ychitre@uckbv5.ece.uc.edu

File Information:

All images are 160 microns and 256 gray levels.

The full images are either 512x512 or 512x480 pixels.

Sub-regions were obtained using binary masks. These regions, corresponding classification of the lesion, and their file size are listed on the following pages.

BENIGN CASES

t4158a	182	166
c2259a	129	193
c2259b	172	193
d6171a	217	201
d6171b	115	141
e8012a	161	168
f10663a	175	176
f13416a	176	183
f13416b	160	171
f13687a	253	203
f13687b	228	289
f14718a	336	289
f14718b	311	320
f15229a	131	151
f15229b	162	161
f15925a	219	71
f15925b	193	129
f19686a	149	137
f19686b	165	225
f20621a	131	106
f20621b	175	170
f21019a	137	116
f21019b	117	146
f23234a	156	129
f23234b	141	146
d8453a	193	182
d8453b	184	197
c4888a	187	175
c4888b	225	225
e5810a	33	52
e5810b	33	44
e2745a	129	141
e2745b	130	65
d6928a	103	129
d6928b	133	141
d9569a	445	385
d9569b	146	118
d3253a	244	129
d3253b	262	172
d7777a	161	169
d7777b	152	148
e0648a	108	109
e0648b	130	111
t2331a	97	97
t2331b	139	129
d7444a	257	270
d7444b	257	239
e9374a	65	65
e9374b	65	97
e7759a	112	134
e7759b	81	91
e9579a	161	161
e9579b	129	161
c9624a	118	112
c9624b	117	129
e7309a	201	215
e7309b	198	143
f15889a	54	50
f15889b	83	101
f14413a	130	135
f14413b	142	184
f15512a	410	299
f15512b	346	282

MALIGNANT

t2186a	161	225
t2186b	161	355
t3060a	245	161
t3060b	148	293
c6271a	353	235
c6271b	225	390
d4194a	208	257
d4194b	231	270
d8027a	153	111
d8027b	94	95
d8027c	162	117
d8027d	129	83
d8027e	129	99
d9834a	110	108
d9834b	74	73
e5454a	232	147
e5454b	193	242
e7916a	353	289
e8030a	310	394
e8030b	437	463
e9986a	97	161
e9986b	237	151
es1552a	152	138
es1552b	104	92
f10020a	129	129
f10020b	129	129
f19791a	229	289
f19791b	225	231
f26930a	177	186
f26930b	170	170
f23259a	181	186

APPENDIX D. Other Databases

Georgetown:

Matthew T. Freedman, MD, MBA
Associate Professor
Clinical Director of IMAC
Georgetown University Medical Center
3800 Reservoir Road NW
Washington, DC 20007-2197
phone: (202) 784-3415
fax: (202) 784-4899

University of South Florida:

Maria Kallergi, PhD.
Department of Radiology
University of South Florida
12901 Bruce B. Downs Blvd., Box 17
Tampa, FL 33612-4799
phone: (813) 975-7873
fax: (813) 975-7831
email: kallergi@rad.udf.edu

Appendix E. Khoros.

Khoros is an open environment for information processing, data visualization, and software development. Its benefits include establishing a repository for application-specific software, encouraging project collaboration and code re-use, and standardization. The common data file format is VIFF. Khoros is actually the name of the image processing package, while Cantata is the visual language.

Setup

To run Khoros, you must have a path to Khoros set up on your SUN workstation. The easiest way to do this is to find someone who already has their .cshrc file edited to include Khoros in the path. You will need their username. Once logged onto the a SUN workstation, type these commands at the prompt (represented by the >):

>cp ~username/.cshrc .

This will copy and paste (cp) the prescribed file to your directory.

>more .cshrc

You can now view .cshrc to make sure that ~/.khoros_env is in the file.

>cp ~username/.Khoros .

You have now copied and pasted .Khoros . to your directory.

OPTIONAL:

>cp ~username/.Xdefaults

Copy/pasted the file that establishes screen colors, etc.

>more .xdefaults

To make sure that .xdefaults has been copied.

- Exit window and log back in.

This will act like an autoexec file and set up all paths needed. Note: You can not use >source .cshrc if you copies .xdefaults because the file is too long.

>cantata

You are now ready to run the program. Keep in mind this program takes a few minutes to boot up.

Khoros Basics and A First Exercise

To learn the basics, it is best to call up an existing image. Without doing any image processing, you can display a image with the following few steps:

1) From the pull-down menus select INPUT SOURCE, then IMAGES. Make your selection (feath.xv), then click on GLYPH. This will create a box ("glyph") that you can drag and drop where you wish.

2) Select OUTPUT-DISPLAY IMAGE. By choosing INTERACTIVE here you will be able to edit (i.e. change the gray-level thresholds) the image that is displayed. Click GLYPH and place your box.

3) Connect these two glyphs by clicking the arrow on the input box to the arrow on the output box.

4) There are two ways to activate the network you just established. Click RUN to have Khoros activate each glyph in sequence, or click the on/off icon in each glyph to run just that command.

5) Move the image by clicking on the top blue bar and dragging.

Glyph Box Format.

The top left icon on the glyph box (bomb) will delete the glyph when clicked on. The center icon (page layout) will bring up the commands that were used to create that glyph. Use the top right icon (on/off) to run or terminate running of the glyph commands.

Several features to note:

- To delete a connection, click on the connection, then select delete connection. You must click on the words.

- If a frown face appears in your glyph box, click on it to read the error message. You can either quit, minimize ( ), or glyph this message.

- If the screen locks up, minimize the work area to see the message in the command tool.

Continue with the exercise by selecting IMAGE PROCESSING, choosing any of these features, then glyph. Make the appropriate connections to place this between the input and output glyphs, then RUN.

Displaying a Mammogram

The instructions below establish a Khoros network specifically designed to input a mammographic image, manipulate this file, and display the image. Figure E.1 illustrates the final Khoros workspace.

1) From the pull-down menus select INPUT SOURCE-INPUT DATA FILE. Choose USER DEFINED and click on USER SPECIFIED FILE. Either type the filename or double click on USER SPECIFIED FILE to list the files in the directory. Select a file and glyph.

2) Choose CONVERSION-RAW FILE FORMAT. Select RAW TO VIFF and for OUTPUT DATA TYPE choose SHORT. For RAW INPUT DATA MACHINE, select VAX. Be sure to put in the appropriate file size for ROW and COLUMN. Glyph.

3) Select ARITHMETIC-LOGICAL OPERATIONS. Click on RIGHT SHIFT and input a value for DISTANCE (2 bits for 1024x1024; 3 bits for 512x512). Glyph.

4) Choose CONVERSIONS-DATA CONVERSIONS. Select DATA TYPE CONVERT and input a 1 for SCALE. Ensure OUTPUT DATA TYPE is BYTE. Glyph.

5) Select ARTHMETIC-UNARY-INVERT-GLYPH.

Choose 6a) or 6b):

6a) OUTPUT-DISPLAY IMAGE-INTERACTIVE-GLYPH.

6b) OUTPUT-PRINT IMAGE-GLYPH

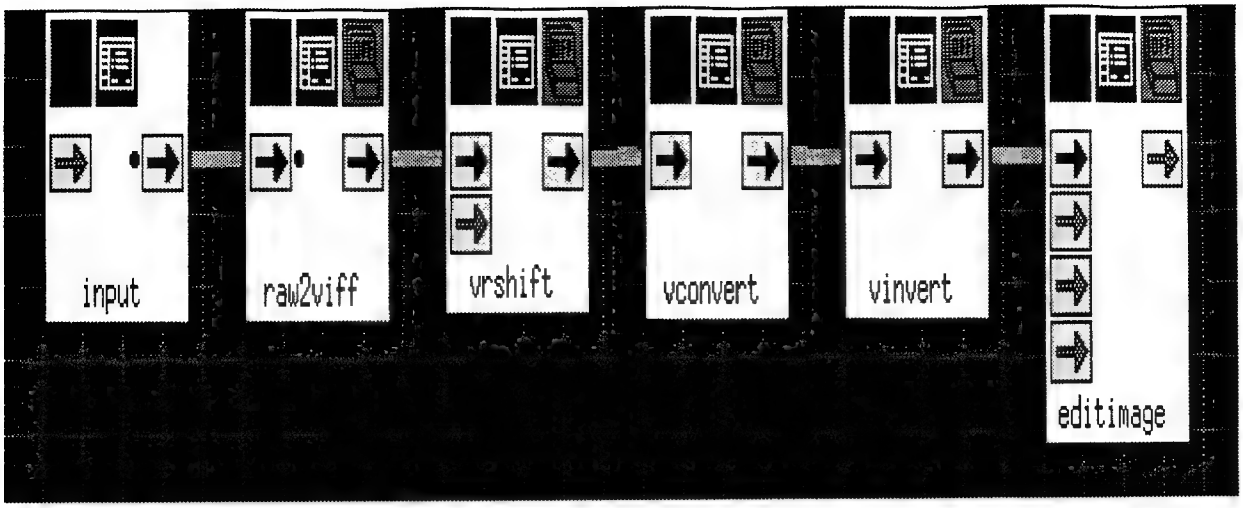


Figure E.1 Khoros Workspace

Histograms

To use the histogram manipulation feature in Khoros:

- Select INPUT SOURCE-INPUT DATA FILE. Choose your image or a user defined file, then glyph.

- Create the next glyph with IMAGE PROCESSING-HISTOGRAMS. Make your choice and GLYPH.

- Select OUTPUT-DISPLAY IMAGE. Choose INTERACTIVE DISPLAY, then glyph.

- Connect the boxes and RUN the network to produce the image.

- DISPLAY UTILITIES shows several ways to manipulate your displayed image.

ZOOM FACTOR: Increasing the factor will increase the zoom magnitude.

ZOOM WINDOW CURSOR: Displays the cursor as a box or cross.

UPDATE MODE: Using continuous will make the displayed area follow the cursor. The button option only updates when clicked.

Printing

NOTE: To use the 600cpi printer in room 243, you must do a remote login to hawkeye before running cantata. To do so, follow these commands:

pinna>rlogin hawkeye (terminal choice is v).

hawkeye>setenv DISPLAY pinna:0.0

(DISPLAY must be in all capital letters and followed by your SUN workstation name--pinna in this example).

pinna>xhost hawkeye

hawkeye>cantata

Commands for Print Glyph

Choose OUTPUT from the pull-down menus, then PRINT IMAGE. For the printer name, leave |pr -P and replace the rest of the line with the appropriate printer:

|pr -Prm243qms 600cpi printer in rm 243

|pr -Psparc Sparc printer.

```
%%%%%  ANGULAR SECOND MOMENT FOR COMPUTER-AIDED BREAST CANCER DIAGNOSIS  %%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  CATHERINE M. KOCUR  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%  AUGUST 1994  %%%%%%%%%%%%%%%%%%%%%%%%%%%%%

clear

%%%%% LOAD ASMVALUES THAT ALREADY EXIST

load asmvalues;

%%%%%  DETERMINE SIZE OF THE INPUT MATRIX

newx = input('ROI name to run asm on:  ','s');
eval(['load ',newx]);
eval(['mtx = ',newx,';']);
[s1 s2]=size(mtx);

%######  take off header info

for i = 1:s2;
  testmtx(i,:) = mtx(i+4,:);
end

[m n] = size(testmtx);


%%%%%  DETERMINE MAXIMUM AND MINIMUN VALUES IN THE MATRIX
%%%%%  THIS IS THE GREY SCALE RANGE FOR THAT IMAGE

MAX = max(max(mtx))
MIN = min(min(mtx))
diff=MAX-MIN

%%%%%  d IS THE GIVEN DISTANCE VECTOR.  THIS STAYS FIXED.


%d = input('input distance vector, format must be [x,y], include brackets  ');
d = [5,0]

%%%%%  COMPUTE THE CO-OCCURRENCE MATRIX
%%%%%  THE VALUE AT EACH LOCATION REPRESENTS HOW MANY TIMES THAT GREY-LEVEL
%%%%%  PAIR OCCURRED AT THE GIVEN DISTANCE VECTOR.
asmom = 0;
count = 0;

number = fix((diff+1)/8);
remain = rem(diff+1,8);
yoffset = number + MIN;

for loop = 1:8,

 if loop == 8,
   occur=zeros(number+remain,diff + 1);
 else
   occur = zeros(number, diff + 1);
 end % end if else

for i=1:m-d(2);
  for j=1:m-d(1);

    a = mtx(i,j);
    b = mtx(i+d(2), j+d(1));
```

```
  if (loop ~= 8) & (a >= number*(loop-1) + MIN) & (a < number * (loop) + MIN)
occur(a-number*(loop-1)-MIN+1,b-MIN+1)=occur(a-number*(loop-1)-MIN+1,b-MIN+1)+1;

  elseif (loop == 8) & (a >= number*(loop-1)+MIN) & (a < number*loop + remain+MIN)
occur(a-number*(loop-1)-MIN+1,b-MIN+1)=occur(a-number*(loop-1)-MIN+1,b-MIN+1)+1;

  end %end if else

  end  % end for j
end  % end for i

asmom = asmom + sum(sum(occur.^2));
count = count + sum(sum(occur));
end % end for loop


%%%%% PRINT OUT OCCURRENCE MATRIX

count
asmom

asval = [asmvalues;0 asmom]

%%%%%  SAVE INTO FILE

save asmvalues asval -ascii
```

```
%########################## Catherine M. Kocur ############################
%################################ Thesis ##################################
%########################## EIGENMASSTRAIN.M ##############################
%################################  1994 ###################################

clear

%#####  set parameters
%#####  m is number of ROI images in training set

m = input('How many training ROIs?  ');

%#####  read in files
%#####  s = number of rows in roi
%#####  t = number of columns in roi
%#####  mtx = the input rois ( 1 thru m)


for r = 1:m;
  fname = input('ROI filename:  ','s');
  eval(['load ',fname]);
  eval(['mtrx' int2str(r) ' = ',fname,';']);
end   %end for r

[e f] = size(mtrx1);

%#####  take off header info

for r = 1:m;
  for i = 1:f;
    eval(['mtx',int2str(r),'(i,:) = mtrx',int2str(r),'(i+4,:);']);
  end
end

[s t] = size(mtx1)

%#####  CREATE XTEMP  --  COLUMN MATRICES  #####
%#####  create x1 matrix with all values in ROI listed in column vector
%##### values are put in across the columns before moving onto next row

for r = 1:m;

k=1;
for i=1:s;
  for j=1:s;
    eval(['xtemp',int2str(r),'(k) = mtx',int2str(r),'(i,j);']);
    k=k+1;
  end  % end for j
end  % end for i

end  % end for r

%##### Transpose all vectors

for r = 1:m;
  eval(['dummy = xtemp',int2str(r),';']);
  dummy = dummy';
  eval(['xtemp',int2str(r),' = dummy',';']);
end


%#####  CALCULATING THE AVERAGEMASS  #####
%#####  AVERAGE OF EACH ELEMENT (PIXEL) IN ALL X VECTORS  #####

sumx = zeros(s^2,1);
for r = 1:m;
```

```
  eval(['sumx = sumx + xtemp',int2str(r),';']);
end

avgx = sumx/m;


%######  NEW x VECTORS ARE THE OLD VECTOR MINUS THE AVERAGE X #####

for r =1:m;
  eval(['x',int2str(r),' = xtemp',int2str(r),' - avgx;']);
end

%######  Create X by augmenting all x1,x2,... together #####

X = x1;
for r = 2:m;
  eval(['X = [X,x',int2str(r),'];']);
end


%######  DETERMINE THE COVARIANCE MATRIX   #####

cv=X'*X;
cov=(1/m)*cv;


%######  FIND THE EIGENVALUES  #####
%######  d is diagonal matrix of eigenvalues  #####
%######  eigen is full matrix whose columns are the corresponding eigenvectors
%######  eigenvectors are scaled so that the norm of each is 1.0

[eigen d] = eig(cov);

%######  CREATE THE EIGENMASS  #####
%######  THESE EIGENMASS REPRESENT ALL INPUT ROIS LOOK WHEN REPRESENTED  #####
%######  TOGETHER  #####

Y = X * eigen;


%######  SEPERATE Y INTO SQUARE MATRICES  #####
%######  COLUMN ONE OF Y IS 1 MTX, COL 2 ANOTHER  #####

for r = 1:m;
  eval(['y',int2str(r),' = Y(:,r);']);
end

%######  CHECK THAT EIGENVECTORS ARE ORTHOGONAL  #####

ortho = y1'*y2*y3'*y4*y5'*y6*y7'*y8*y9'*y10

%##### SAVE FILES

save Y Y -ascii
save eigenvector eigen -ascii
save eigenvalues d -ascii

%######################  END OF PROGRAM  ####################################
```

```
%########################## Catherine M. Kocur ##########################
%############################### Thesis ################################
%############################ EIGENTEST.M ##############################
%############################### 1994 ##################################

clear
load Y

%#####  m is number of ROI images in training set

%m = input('How many ROIs where used to make the training eigenmass?  ');
m=10;

%#####  SEPERATE Y INTO SQUARE MATRICES  #####
%#####  COLUMN ONE OF Y IS 1 MTX, COL 2 ANOTHER  #####

for r = 1:m;
  eval(['y',int2str(r),' = Y(:,r);']);
end

%#####  CALCULATE B COEFF BY DOT PRODUCT OF A NEW X (ROI) AND #####
%#####  Y1 & Y2 (EIGENMASSES)  #####

%##### READ IN ROI TO CALCULATE COEFF

newx = input('ROI name to test:  ','s');
eval(['load ',newx]);
eval(['testmtx1 = ',newx,';']);
[s1 s2] = size(testmtx1);

%#####  take off header info

for i = 1:s2;
  testmtx(i,:) = testmtx1(i+4,:);
end

[a b] = size(testmtx)


%#####  PUT INTO COLUMN FORM

k = 1;
for i = 1:a;
  for j=1:a;
    test(k) = testmtx(i,j);
    k = k + 1;
  end  % end for j
end  % end for i

%##### CALCULATE COEFFICIENTS

for r = 1:m;
  eval(['b',int2str(r),' = test * y',int2str(r),';']);
end

bcoeff = [b1 b2 b3 b4 b5 b6 b7 b8 b9 b10]

save bcoeff bcoeff -ascii

%######################  END OF PROGRAM  ######################################
```

```
clear

fprintf('\n')
fprintf('\n')
fprintf('****                    SQROI.M                          ****\n')
fprintf('****                  THIS PROGRAM                       ****\n')
fprintf('****        CREATES A SQUARE ROI FROM A SUBREGION        ****\n')
fprintf('****            Written by Catherine Kocur               ****\n')
fprintf('****                    October 1994                     ****\n')
fprintf('\n')
fprintf('****  NOTE:  Be sure to change the value that is hard    ****\n'
fprintf('****         coded for benign or malignant               ****\n')
fprintf('\n')
fprintf('\n')

load d8027e.ascii;
mtx = d8027e;

%%%%%%%  a = # cols, b = # rows

[a b] = size(mtx)
a = a - 4;

%%%%%%%%  find center point

rowcenter = floor(a/2);
colcenter = floor(b/2);

%%%  create square roi around center point

startcol = colcenter-15;
endcol = colcenter+16;
startrow = rowcenter-15+4;
endrow = rowcenter+16+4;

m=1;
n=1;
for i=startrow:endrow;
  for j=startcol:endcol;

  new(m,n)=mtx(i,j);
  n=n+1;

  end % for j

  m=m+1;
  n=1;
end %for i

%%%%  roi info
head = zeros(4,32);

%%%%%%  malignant=109   benign=98
head(1,1) = 109;
head(2,1) = 32;
head(3,1) = 32;
head(4,1) = 1;

%%%%%%  augment the header and the new roi

roi = [head;new]

%%%%%% save as a new file

save roid8027e roi -ascii
```

Bibliography

1. The American Heritage Dictionary (Second College Edition). Boston: Houghton Mifflin Company, 1982.

2. Bankman, Issac, William Christens-Barry, Dong Kim, Irving Weinberg, Olga Gatewood, and William Brody. "Automated Recognition of Microcalcification Clusters in Mammograms," SPIE, v1905: 731-737 (1993).

3. Benoit-Cattin, Hugues, Olivier Baudin, Atilla Baskurt, and Robert Goutte. "Coding Mammograms Using Wavelet Transform," SPIE, v 2164: 282-290 (1994).

4. Boas, Mary L. Mathematical Methods in the Physical Sciences (Second Edition). New York: John Wiley & Sons, Inc., 1983.

5. Burns, Rogers, Oxley, Ruck. "Motion Segmentation Using a Non-Homogeneous L_2R^3 Multiresolution Wavelet Analysis," Submitted to IEEE Trans. On Aerospace and Electronic Systems, October 1993.

6. Caldwell, Curtis, Sandra Stapleton, David Holdsworth, Roberta Jong, William Weiser, Gabriel Cooke, and Martin Yaffe. "Characterization of Mammographic Parenchymal Pattern by Fractal Dimension," Phys. Med. Biol., v35: 235-247 (1990).

7. Chan HP, K Doi, S Galhotra, Vyborny CJ, H MacMahon, and PM Jokich. "Image Feature Analysis and Computer-Aided Diagnosis in Digital Radiography: Automated Detection of Microcalcifications in Mammography," Medical Physics, v14: 538-548 (1987).

8. Chang, Tianhorng, and C.-C. Jay Kuo. "Texture Analysis and Classification with Tree-Structured Wavelet Transform," IEEE Transaction on Image Processing, v2: 429-441 (October 1993).

9. Chitre, Yateen, Atam P. Dhawan, and Myron Moskowitz. "Artificial Neural Network Based Classification of Mammographic Microcalcifications Using Image Structure Features," International Journal of Pattern Recognition and Artificial Intelligence, v7: 1377-1401.

10. Chitre, Yateen, Atam Dhawan, and Myron Moskowitz. "Artificial Neural Network Based Classification of Mammographic Microcalcifications Using Image Structure Features," IEEE Engineering in Medicine and Biology, v15: 50-51 (1993).

11. Cybenko, G. "Approximation by Superpositions of a Sigmoidal Function," Math. Control, Signals Sys., v2: 303-314 (1989).

12. Dhawan, Atam P. "Computerized Mammographic Image Analysis for Reducing False Positive Rate for Biopsy Recommendation," SPIE, v1905: 540-541 (1993).

13. Dhawan, Atam, Yateen Chitre, and Myron Moskowitz. "Artificial Neural Network Based Classification of Mammographic Microcalcifications Using Image Structure Features," SPIE, v1905: 820-831 (1993).

14. Dhawan, Atam, Yateen Chitre, Myron Moskowitz, and Eric Gruenstein. "Classification of Mammographic Microcalcification and Structural Features Using An Artificial Neural Network," IEEE Engineering in Medicine and Biology, v13(3): 1105-1106 (1991).

15. EK. "Hubble CCD Used for Biopsies," Optics & Photonics News: (October 1994).

16. Eisenbies, Christopher. Classification of Ultra High Resolution Radar Using Decision Boundary Analysis. MS Thesis. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1994.

17. Giger, Maryellen. "Computer-Aided Diagnosis," RSNA Categorical Course in Physics: 283-298 (1993).

18. Hajnal, Saki, Paul Taylor, Marie-Helene Dilhuydy, Beatrice Barreau and John Fox. "Classifying Mammograms by Density: Rationale and Preliminary Results," SPIE, v1905: 478-489 (1993).

19. Harrup, Georgia. ROC Analysis of IR Segmentation Techniques. MS Thesis. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1994.

20. Haus, Arthur G. "Screen-Film Image Receptors and Film Processing," RSNA Categorical Course in Physics: 83-99 (1993).

21. Haus, Arthur, and Martin Yaffe (Editors). Syllabus: A Categorical Course in Physics - Technical Aspects of Breast Imaging (Second Edition). Oak Brook: RSNA, 1993.

22. Hendrick, R. Edward, and Steve Parker. "Stereotaxic Imaging," RSNA Categorical Course in Physics: 257-269 (1993).

23. Hoffmeister, Jeffery W. Aerospace Physician. AL/CFHO, Wright-Patterson AFB, OH. Personal interviews. May - November 1994.

24. Hush, Don and Bill Horne. "Progress in Supervised Neural Networks: What's New Since Lippmann," IEEE Signal Processing Magazine: 8-39 (January 1993).

25. Kelley, John L. Development of a Mammographic Image Processing Environment Using Matlab. MS Thesis. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1994.

26. Krupp, Marcus, Steven Schroeder, and Lawrence Tierney. Current Medical Diagnosis & Treatment 1987. Norwalk: Appleton & Lange, 1987.

27. Kukolich, Linda, and Richard Lippmann. LNKnet User's Guide. MIT Lincoln Laboratory, July 1993.

28. Laine, Andrew, Sergio Schuler, Walter Huda, Janice Honeyman, and Barbara Steinbach. "Hexagonal Wavelet Processing of Digital Mammography," SPIE Image Processing, v1898: 559-573 (1993).

29. Laine, Andrew, and Shuwu Song. "Multiscale Wavelet Representations for Mammographic Feature Analysis," SPIE Mathematical Methods in Medical Imaging, v1768: 306-316 (1992).

30. Lee, Chulhee, and David Langrebe. "Feature Extraction Based on Decision Boundaries," IEEE Transactions on Pattern Analysis and Machine Intelligence, v15(4): 388-400 (April 1993).

31. Mayo Foundation for Medical Education and Research. Breast Cancer: New Perspectives Can Replace Unrealistic Fears. Rochester, MN: Supplement to Mayo Clinic Health Letter, October 1994 (ISSN 0741-6245).

32. Metz, Charles. "ROC Methodology in Radiologic Imaging," Investigative Radiology, v21: 720-733 (September 1986).

33. Miller, Peter, and Sue Astley. "Classification of Breast Tissue by Texture Analysis," Image and Vision Computing, v10: 277-281 (June 1992).

34. Myers, Lemuel Jr. Image Interpretation and Enhancement for the Visually Impaired. MS Thesis. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1994.

35. Nishikawa, Robert, Maryellen Giger, Kunio Doi, Charles Metz, Fang-Fang Yin, Carl Vyborny, and Robert Schmidt. "Effect of Case Selection on the Performance of Computer-Aided Detection Schemes," Medical Physics, v21: 265-269 (Feb 1994).

36. Pietikainen, Matti, Azriel Rosenfeld, and Larry Davis. "Texture Classification Using Averages of Local Pattern Matches," IEEE Conference on Pattern Recognition, v1: 301-302 (Oct 1982).

37. Priddy, Kevin, Steven Rogers, Dennis Ruck, Gregory Tarr, and Matthew Kabrisky. "Bayesian Selection of Important Features for Feedforward Neural Networks," Neurocomputing, v5: 91-103 (1993).

38. Rogers, Steven, Matthew Kabrisky, Dennis Ruck, and Gregory Tarr. An Introduction to Biological and Artificial Neural Networks. City: Publisher, 1990.

39. Rogers, Steven, Dennis Ruck, and Matthew Kabrisky. "Artificial Neural Networks for Early Detection and Diagnosis of Cancer," Cancer Letters, v77: 79-83 (1994).

40. Rothnberg, Mikel, and Charles Chapman. Dictionary of Medical Terms (Second Edition). Hauppauge: Barron's Educational Services, Inc., 1989.

41. Ruck, Dennis, Steven Rogers, and Matthew Kabrisky. "Feature Selection Using a Multilayer Perceptron," Journal of Neural Network Computing, v2: 40-48 (1990).

42. Ruck, Dennis, Steven Rogers, Matthew Kabrisky, Mark Oxley, and Bruce Suter. "The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function," IEEE Transactions on Neural Networks, v1: 296-298 (Dec 1990).

43. Sanders, Mark and Ernest McCormick. Human Factors in Engineering and Design (Seventh Edition). New York: McGraw-Hill, Inc., 1993.

44. Sarwal, Alok, Yateen Chitre, and Atam Dhawan. "Segmentation of Mammographic Microcalcifications," IEEE Engineering in Medicine and Biology, v15: 112-113 (1993).

45. Schalkoff, Robert. Digital Image Processing and Computer Vision. New York: John Wiley & Sons, Inc., 1989.

46. Schalkoff, Robert. Pattern Recognition: Statistical, Structural, and Neural Approaches. New York: John Wiley & Sons, Inc., 1992.

47. Shaw de Paredes, Ellen. "Radiologic Breast Anatomy: Radiologic Signs of Breast Cancer," RSNA Categorical Course in Physics: 35-46 (1993).

48. Silverberg, E., and J. Lubera. "Cancer Statistics," Cancer, v39: (1987).

49. Singstock, Brain. Infrared Target Recognition. MS Thesis. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1991.

50. Steppe, Jean M. Feature and Model Selection in Feedforward Neural Networks. PhD dissertation. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, June 1994.

51. Steppe, Bauer, and Rogers. "Integrated Feature and Architecture Selection," Submitted to IEEE Trans. on Neural Networks, 1994.

52. Suarez, Pedro F. Face Recognition with the Karhunen-Loeve Transform. MS thesis. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1991.

53. Tanne, Janice Hopkins. "Everything You Need to Know About Breast Cancer...But Were Afraid to Ask," New York [GNYC], v26: 52-62 (Oct 1993).

54. Tarr, G.L. Multi-Layer Feedforward Neural Networks for Image Segmentation. PhD Dissertation. Wright-Patterson AFB, Ohio: School Of Engineering, Air Force Institute of Technology, December 1991.

55. Taveras, Juan, and Joseph Ferrucci (Editors). Radiology: Diagnosis-Imaging-Intervention (Volume 1). Philadelphia: J.B. Lippincott Company, 1990.

56. Vyborny, Carl, and Robert Schmidt. "Technical Image Quality and the Visibility of Mammographic Detail," RSNA Categorical Course in Physics: 101-109 (1993).

57. Wu, Yuzheng, Maryellen Giger, Kunio Doi, Carl Vyborny, Robert Schmidt, and Charles Metz. "Artificial Neural Networks in Mammography: Application to Decision Making in the Diagnosis of Breast Cancer," Radiology, v187: 81-87 (April 1993).

58. Yin, Fang-Fang, Maryellen Giger, Kunio Doi, Charles Metz, Carl Vyborny, and Robert Schmidt. "Computerized Detection of Masses in Digital Mammograms: Analysis of Bilateral Subtraction Images," Medical Physics, v18: 955-963 (Sep/Oct 1991).

59. Yaffe, Marin J. "Digital Mammography," RSNA Categorical Course in Physics: 271-282 (1993).

60. Yoshida, Hiroyuki, Kunio Doi, and Robert Nishikawa. "Automated Detection of Clustered Microcalcifications in Digital Mammograms Using Wavelet Transform Techniques," SPIE Image Processing, v2167: 868-886 (1994).

Vita

Captain Catherine M. Kocur was born on 7 March 1967 in Pleasantville, New York. She graduated from Pleasantville High School in 1985 and attended the United States Air Force Academy, graduating with a Bachelor of Science in Space Operations in May 1989. Upon graduation, she received a regular commission in the United States Air Force. After completing Undergraduate Space Training at Lowry AFB, Colorado, she served her first tour of duty at Eglin AFB, Florida. She began as a Space Operation Crew Commander responsible for a crew which managed the AN/FPS-85 Phased Array Radar and its spacetrack mission of satellite tracking and data collection on high interest space activity. She was later chosen for various position including Chief, Operations Training, Chief, Standardization/Evaluation, and commander's Executive Officer. Captain Kocur entered the School of Engineering, Air Force Institute of Technology, in May 1993, to obtain her Masters of Science degree in Space Operations, with concentration in Pattern Recognition.

Permanent Address:

168 Bedford Road
Pleasantville, New York 10570

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1994	Master's Thesis

4. TITLE AND SUBTITLE

COMPUTER-AIDED BREAST CANCER DIAGNOSIS

5. FUNDING NUMBERS

6. AUTHOR(S)

Catherine M. Kocur, Captain, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GSO/ENG/94D-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Jeffrey W. Hoffmeister, Captain, USAF
AL/CFHD
2255 H. Street, Bldg 248A
WPAFB OH 45433-7022

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

To be published in IEEE Engineering in Medicine and Biology

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This research advances computer-aided breast cancer diagnosis. More than 50 million women over the age of 40 are currently at risk from this disease in the United States. Computer-aided diagnosis is offered as a *second opinion* to radiologists to aid in decreasing the number of false readings of mammograms. This automated tool is designed to enhance detection and classification.

New feature extraction methods are presented that provide increased classification power. Angular second moment, a second-order gray-level histogram statistic, provides baseline accuracy. Two novel extraction methods, *eigenmass* and wavelets, are introduced to the field. Based on the Karhunen-Loeve Transform, *eigenmass* features are developed using eigenvectors to alter the data set into new coefficients. Wavelets, previously only exploited for their segmentation benefits, are explored as features for classification. Daubechies_4, Daubechies_20, and biorthogonal wavelets are each investigated. Applied to 94 *difficult-to-diagnosis* digitized microcalcification cases, performance is 74 percent correct classifications.

Feature selection techniques are presented which further improve performance. Statistical analysis, neural and decision boundary-based methods are implemented, compared, and validated to isolate and remove useless features. The contribution from this analysis is an increase to 88 percent correct classification in system performance.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Computer-Aided Diagnosis, neural network, breast cancer, feature selection, Karhunen-Loeve Transform, wavelet	109
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL